

# False Discovery Rate Control via Debiased Lasso

Adel Javanmard\*

Data Sciences and Operations Department, University of Southern California  
e-mail: [ajavanma@usc.edu](mailto:ajavanma@usc.edu)

Hamid Javadi

Department of Electrical and Computer Engineering, Rice University  
e-mail: [hrhakim@rice.edu](mailto:hrhakim@rice.edu)

## Abstract:

We consider the problem of variable selection in high-dimensional statistical models where the goal is to report a set of variables, out of many predictors  $X_1, \dots, X_p$ , that are relevant to a response of interest. For linear high-dimensional model, where the number of parameters exceeds the number of samples ( $p > n$ ), we propose a procedure for variables selection and prove that it controls the *directional* false discovery rate (FDR) below a pre-assigned significance level  $q \in [0, 1]$ . We further analyze the statistical power of our framework and show that for designs with subgaussian rows and a common precision matrix  $\Omega \in \mathbb{R}^{p \times p}$ , if the minimum nonzero parameter  $\theta_{\min}$  satisfies

$$\sqrt{n}\theta_{\min} - \sigma \sqrt{2(\max_{i \in [p]} \Omega_{ii}) \log \left( \frac{2p}{qs_0} \right)} \rightarrow \infty,$$

then this procedure achieves asymptotic power one.

Our framework is built upon the debiasing approach and assumes the standard condition  $s_0 = o(\sqrt{n}/(\log p)^2)$ , where  $s_0$  indicates the number of true positives among the  $p$  features. Notably, this framework achieves exact directional FDR control without any assumption on the amplitude of unknown regression parameters, and does not require any knowledge of the distribution of covariates or the noise level. We test our method in synthetic and real data experiments to assess its performance and to corroborate our theoretical results.

**MSC 2010 subject classifications:** Primary 62F03, 62J05, 62J07; secondary 62F12.

**Keywords and phrases:** Inference in high-dimensional regression, Hypothesis testing, False discovery rate, Model selection, Lasso, Debiased estimator.

## Contents

1	Introduction . . . . .	1
---	------------------------	---

\*A. Javanmard is supported in part by NSF CAREER Award 1844481 and a Google Faculty Research Award. A. Javanmard would also like to acknowledge the financial support of the Office of the Provost at the University of Southern California through the Zumberge Fund Individual Grant Program.

1.1	Problem Formulation . . . . .	3
1.2	Our Contributions and Outline of the Paper . . . . .	5
1.3	Further Related Work . . . . .	6
1.4	Notations . . . . .	7
2	FCD Procedure: False Discovery Control via Debiasing . . . . .	8
2.1	Debiasing Lasso . . . . .	8
2.2	Extension to Non-Gaussian Noise . . . . .	10
2.3	FCD Procedure . . . . .	11
2.3.1	Construction of Test Statistics . . . . .	11
2.3.2	A Data Dependent Threshold for the Test Statistics . . . . .	12
3	Main Results . . . . .	12
3.1	Control of Directional False Discovery Rate . . . . .	12
3.2	Power Analysis . . . . .	13
4	Improved Results for Gaussian Designs . . . . .	14
4.1	Known Covariance . . . . .	15
4.2	Unknown Covariance . . . . .	16
5	Numerical Experiments . . . . .	17
6	Real Data Experiments . . . . .	19
7	Proof of Main Theorems . . . . .	20
7.1	Proof of Theorem 3.1 . . . . .	20
7.1.1	Highly correlated test statistics $(\Gamma(\gamma, c_0))$ . . . . .	25
7.1.2	Weakly correlated test statistics $(\Gamma(\gamma, c_0)^c)$ . . . . .	28
7.2	Proof of Theorem 3.3 . . . . .	30
7.3	Proof of Theorem 4.1 . . . . .	32
7.4	Proof of Theorem 4.2 . . . . .	33
	Acknowledgements . . . . .	34
A	Proof of Technical Lemmas . . . . .	35
A.1	Proof of Lemma 7.2 . . . . .	35
A.2	Proof of Lemma 7.3 . . . . .	35
A.3	Proof of Lemma 7.4 . . . . .	36
A.4	Proof of Corollary 3.5 . . . . .	38
	References . . . . .	38

## 1. Introduction

Living in the era of data deluge, modern datasets are often very fine-grained, including information on a large number of potential explanatory variables. For a given response of interest, we know a priori that a large portion of these variables are irrelevant and would like to select a set of predictors that influence the response. For example, in genome-wide association studies (GWAS), we collect single nucleotide polymorphism (SNP) information across a large number of loci and then aim at finding loci that are related to the trait, while being resilient to false associations.

The focus of this paper is on high-dimensional regression models where the number of parameters exceeds the sample size. Since such models are over-parameterized, they are prone to overfitting. In addition, high-dimensionality brings noise accumulation and spurious correlations between response and unrelated features, which may lead to wrong statistical inference and false predictions. Model selection is therefore a crucial task in analyzing high-dimensional models. For a successful model selection, we need to assure that most of the selected predictors are indeed relevant. This not only leads to noise reduction and enhances predictions but also offers reproducibility.

To be concrete in using the term “reproducibility”, we characterize it for statistical inference problem by using the false discovery rate (FDR) criterion, which is the expected fraction of discoveries that are false positives. The notion of FDR has been proposed by the groundbreaking work [BH95] and nowadays is the criterion of choice for statistical inference in large scale hypothesis testing problem. In their work, Benjamini and Hochberg developed a procedure to control FDR under a pre-assigned significance level. It has been shown theoretically that BH procedure controls FDR in some special cases such as independence or positive dependence of tests [BH95, BY01]. Since initially proposed, there have been various modifications of BH [BY01, SRC<sup>+</sup>15, FHG12, Wu08, XCML11] and its applications in different domains [RYB03, GLN02].

Importantly, BH procedure (and its modifications) assumes that  $p$ -values are given as input for all the hypothesis tests. The  $p$ -values are often calculated using classical formula obtained by using large-sample theory which are theoretically justified only for the classical setting of fixed dimension  $p$  and diverging sample size  $n$  [VdV00]. For example, [LS<sup>+</sup>14] considers the setting where  $m$  i.i.d random samples of  $(X_1, \dots, X_p)$  are given and  $p$ -values are estimated from the Student’s  $t$ -test statistic. The authors propose a bootstrap calibration method to use with the BH procedure and show that, under weak dependence among observations, it can control the false discovery rate when the total number of observations  $n = mp$  is bigger than  $p(\log p)^c$ . However, for high-dimensional models obtaining valid  $p$ -values is highly nontrivial. This is in part due to the fact that fitting high-dimensional model often requires the use of nonlinear and non-explicit estimation procedures and characterizing the distribution of such estimators is extremely challenging. In the past couple years, there has been a surge of interest in constructing frequentist  $p$ -values and confidence intervals for high-dimensional models. A common approach is the fundamental idea of debiasing which was proposed in a series of work [JM14b, ZZ14, JM14a, VdGBRD14, JM13b, BCH14]. In this approach, starting from a regularized estimator one first constructs a debiased estimator and then makes inference based on the asymptotic normality of low-dimensional functionals of the debiased estimator. This approach also provides asymptotically valid  $p$ -values for the null hypotheses of the form  $H_0 : \theta_{0,i} = 0$ , where  $\theta_{0,i}$  is a fixed single model parameter. However, these  $p$ -values are correlated and the BH procedure is not guaranteed to control FDR in this case. The modification of BH for general dependency, that scales the significance level by  $1/(\log p)$  factor [BY01], also turns out to be overly conservative and leads to a low power. In [BCC<sup>+</sup>18], the authors review the

methods for constructing  $p$ -values in the high dimensional setting, their behavior and limitations, and describe a general set of assumptions under which these  $p$ -values can be used for inference tasks, such as finding confidence intervals and controlling FDR. In particular, [BCC<sup>+</sup>18] extends the result of [LS<sup>+</sup>14] to the so-called “Many Approximation Means (MAM)” framework and provide a set of conditions on the dependence among  $p$ -values such that the Benjamini-Hochberg procedure has the FDR control property.

In this paper, we build upon the debiasing approach and propose a procedure for model selection under the high-dimensional regime, which is guaranteed to have FDR under a pre-assigned level  $\alpha \in [0, 1]$ . We call our procedure *FCD* (for “*FDR Control via Debiasing*”) and prove that it controls even a stronger criterion, namely *directional* FDR. We further analyze its statistical power (without imposing any assumption on the amplitude of the regression parameters or the noise level).

Controlling FDR in regression model has been a long standing problem. It was just a couple years ago that [BC15] proposed the ingenious idea of knockoff filter. In a nutshell, this approach constructs a set of “knockoff” variables that are irrelevant to response (conditional on the original covariates) but whose structure mirrors that of the original covariates. The knockoff variables then behave as the controls for original covariates. This way, they bypass the need of constructing  $p$ -values and directly select a model with the desired FDR. The focus of [BC15] was on linear regression model with  $n > 2p$ . Later, [CFJL18] extended the idea of knockoff filter to high-dimensional nonlinear models with random designs, but assumes that the joint distribution of covariates is known. Very recently, [FDLL17, BCS18] studied robustness of model-X knockoff to errors in estimating the joint distribution of covariates. A salient feature of the knockoff approach is that for  $n \geq 2p$ , it controls FDR in finite sample setting without requiring any assumption on the covariates. However, the extension model-X knockoffs [CFJL18] requires the knowledge of the joint distribution of covariates. Moreover, the knockoff approach does not provide valid  $p$ -values for the hypotheses regarding the model parameters. By contrast, the FCD method that we present in this paper controls FDR as long as  $s_0 = o(\sqrt{n}/(\log p)^2)$ , without requiring the joint distribution of covariates. Furthermore, it comes with the valid  $p$ -values for individual model parameters. However, the FDR control is proved for the asymptotic regime where  $n \rightarrow \infty$ .<sup>1</sup>

### 1.1. Problem Formulation

Suppose we have recorded  $n$  i.i.d observational units  $(y_1, x_1), \dots, (y_n, x_n)$ , with  $y_i \in \mathbb{R}$  representing response variables and  $x_i \in \mathbb{R}^p$  indicating the vector of explanatory variables on each sample, also referred to as features. We assume the classical linear regression model where the observations obey the following

<sup>1</sup>A finite sample analysis of our method is possible but requires a more involved analysis and is out of the scope of the present work.

relation:

$$y_i = \langle \theta_0, x_i \rangle + w_i, \quad (1)$$

Here,  $\theta_0 \in \mathbb{R}^p$  is the unknown vector of coefficients. The symbol  $\langle \cdot, \cdot \rangle$  denotes the standard inner product. Let  $y = (y_1, \dots, y_n)^\top$  and let  $X \in \mathbb{R}^{n \times p}$  denote the feature matrix that have  $x_1^\top, \dots, x_n^\top$  as rows. Then, writing the linear regression model in matrix form, we obtain

$$y = X\theta_0 + w, \quad (2)$$

We assume that conditional on the design  $X$ , the noise variables  $w_i$  are independent with

$$\mathbb{E}(w_i|X) = 0, \quad \mathbb{E}(w_i^2|X) = \sigma^2, \quad \mathbb{E}(|w_i|^{2+a}|X) \leq C\sigma^{2+a}, \quad (3)$$

for some constants  $C > 0$ ,  $a > 2$ .

We let  $S \subseteq \{1, \dots, p\}$  denote the set of truly relevant feature variables among the many that have been recorded. This set corresponds to the support of  $\theta_0$ , i.e.,

$$S \equiv \text{supp}(\theta_0) = \{1 \leq i \leq p : \theta_{0,i} \neq 0\}. \quad (4)$$

We let  $s_0 = |S|$  be the size of support or in other words the number of true positives.

In this paper, we propose a framework to select a set  $\widehat{S}$  of the feature variables, while controlling the directional false discovery rate (FDR) for the selected variables. This criterion is intimately related to type S errors (S stands for sign). Type S error (a.k.a type III error) occurs when we say, with confidence, that a comparison goes one way while it goes the other way [GT00]. For example, we claim that  $\theta_1 > \theta_2$ , with confidence, while in fact  $\theta_1 < \theta_2$ . In other words, we mistakenly make a claim on the sign (direction) of  $\theta_1 - \theta_2$ . Gelman et. al. [GT00] argue that type S error is a more relevant notion in many applications. Tukey also conveys a similar message in [Tuk91] by arguing that the effects of  $A$  and  $B$ , for any  $A$  and  $B$ , are always different (in some decimal precision) and hence instead of questioning whether there is any difference in two effects, the valid question should be about the direction in which effect of  $A$  differs from that of  $B$ .

Motivated by this, we formally define directional FDR, denoted by  $\text{FDR}_{\text{dir}}$ . For a selected set  $\widehat{S}$  of the features along with estimates  $\widehat{\text{sign}}_j \in \{-1, +1\}$  of the sign of  $\theta_{0,j}$ , we define

$$\text{FDR}_{\text{dir}} = \mathbb{E}[\text{FDP}_{\text{dir}}], \quad \text{FDP}_{\text{dir}} = \frac{|\{j \in \widehat{S} : \widehat{\text{sign}}_j \neq \text{sign}(\theta_{0,j})\}|}{\max(|\widehat{S}|, 1)}, \quad (5)$$

where we adopt the convention  $\text{sign}(0) = 0$ . In words,  $\text{FDR}_{\text{dir}}$  is the expected fraction of false discoveries among the selected ones, where a false discovery is

measured with respect to type S and type I errors. For example, if  $\widehat{\text{sign}}_j = +1$ , while  $\theta_{0,j} = 0$  (type I error) or  $\theta_{0,j} < 0$  (type S error), it is considered as a false discovery.

Recall that the classical FDR is defined as

$$\text{FDR} = \mathbb{E} \left[ \frac{|\{j \in \widehat{S} : \theta_{0,j} \neq 0\}|}{\max(|\widehat{S}|, 1)} \right], \quad (6)$$

that defines the false discoveries only with respect to type I error. Therefore, a comparison of definitions (5) and (6) reveals that

$$\text{FDR}_{\text{dir}} \geq \text{FDR}, \quad (7)$$

for any selected set  $\widehat{S}$ . As a result, proving that a framework controls  $\text{FDR}_{\text{dir}}$  automatically implies that it also controls FDR.

Likewise, we define the statistical power of a selected set  $\widehat{S}$  as

$$\text{Power} = \mathbb{E} \left[ \frac{|\{j \in \widehat{S} : \widehat{\text{sign}}_j = \text{sign}(\theta_{0,j})\}|}{\max(|\widehat{S}|, 1)} \right], \quad (8)$$

i.e., for a true discovery not only the corresponding variable should be in fact non-zero but we should also declare its sign correctly.

The directional FDR has been also studied by [BC16] and it is shown that the knockoff filter also controls this metric as well as the FDR.

## 1.2. Our Contributions and Outline of the Paper

Here, we provide a vignette of our contributions:

**Controlling directional FDR** In Section 2, we propose a method for selecting relevant variables using the debiasing approach. We use the acronym FCD to call this method (standing for “FDR Control via Debiasing”). In Section 3, we show that for design matrices with subgaussian rows, under the standard condition  $s_0 = o(\sqrt{n}/(\log p)^2)$ , the FCD framework achieves exact directional FDR control. (See Theorem 3.1 for a formal statement).

**Characterizing the statistical power** In Section 3.2, we characterize the statistical power of the FCD method. In particular, for designs with subgaussian rows and a common precision matrix  $\Omega \in \mathbb{R}^{p \times p}$ , we show that if the minimum nonzero coefficient,  $\theta_{\min}$  satisfies

$$\sqrt{n}\theta_{\min} - \sigma \sqrt{2(\max_{i \in [p]} \Omega_{ii}) \log \left( \frac{2p}{qs_0} \right)} \rightarrow \infty,$$

then FCD achieves asymptotic power one.

Recently, [FDLL17] has studied the power of model-X knockoff filter, provided that  $\theta_{\min} \sqrt{\frac{n}{\log p}} \rightarrow \infty$  and assuming a lower bound on the size of the

model selected by the knockoff procedure. Namely, if  $|\widehat{S}| \geq cs_0$ , for some constant  $c \in (2(qs_0)^{-1}, 1)$ . Under such assumptions, it is shown that the model-X knockoff approach achieves asymptotic power one. Other than being restrictive, these assumptions are hard to verify and a sufficient given condition is that the size of  $\{j : |\theta_{0j}| \gg \sqrt{s_0(\log p)/n}\}$  is at least  $cs_0$ , for some constant  $c \in (2(qs_0)^{-1}, 1)$ . This condition on the amplitude of nonzero coefficients is much stronger than the one we need for FCD to achieve power one.

**Numerical validation** We validate our approach on both synthetic and real data in Sections 5 and 6 and compare its performance with the model-X knockoff. As the simulations suggest, FCD method compares favorably to the model free knockoff in a wide range of setups. We also compare the statistical power of FCD with the theoretical characterization and show that they are in good agreement.

**Techniques.** In our analysis of FDR, we use ideas from the debiasing approach [JM14b, ZZ14, JM14a, VdGBRD14, JM13b] together with some results from [Liu13] regarding the order statistic of sum of Gaussian random variables (See Lemma 6.1, 6.2 therein.) It is worth mentioning that [Liu13] developed such results to use in the analysis of a method they proposed for Gaussian graphical model and its FDR. This context is very different from the problem studied in this paper and as expected the test statistics are also very different. In our FCD approach, we construct the test statistics by debiasing the Lasso solution. These test statistics have a Gaussian part and a bias term. In applying the results from [Liu13], we need to do a careful analysis of the bias term, and also the errors in noise level estimation. In addition, by a careful analysis of the test statistic and the data dependent threshold used in our procedure, we are able to analyze the statistical power of our approach.

### 1.3. Further Related Work

There exists a copious theoretical literature developed on high-dimensional regression and the Lasso. Most existing studies have focused on prediction error [GR04], model selection properties [MB06, ZY06, Wai09, CP09], estimation consistency [CT05, BRT09]. For exact support recovery, it was known early on that, even in the classical setting of fixed  $p$  and diverging  $n$ , support of Lasso will be different from  $S$  (support of true signal) unless the columns of  $X$ , with index in  $S$ , are roughly orthogonal to the ones with index outside  $S$  [KF00]. This assumption is formalized under the so-called ‘irrepresentability condition’. In a seminal work, Zhao and Yu [ZY06] show that this condition also allows exact support recovery in the high-dimensional setting ( $p \gg n$ ). Independently, [MB06] studied model selection problem for random Gaussian designs, with applications to learning Gaussian graphical models. These papers consider the setting of  $s_0 = O(n^c)$ , for some  $c < 1$ . Further, under a normalization of design such that its columns have norm at most  $\sqrt{n}$ , they require the minimum nonzero amplitude of the signal  $\theta_{\min} = \min_{i \in S} \theta_{0,i}$  to satisfy  $\theta_{\min} > c\sqrt{s_0/n}$ . Later, [Wai09]

improved these results for the case of random Gaussian designs and showed that for a broad range of covariance matrices, the Lasso can recover the support of a signal for which  $\theta_{\min} > c\sigma\sqrt{(\log p)/n}$ . The model selection problem was also studied under the weaker, generalized irrepresentability condition, for the Gauss-Lasso estimator [JM13a].

As an alternative to irrepresentability condition, [Lou08] proves the exact model selection under an incoherence assumption of  $\max_{i \neq j} \widehat{\Sigma}_{ij} = O(1/s_0)$ . This assumption is however stronger than irrepresentability condition [vdGB09].

As discussed in the introduction, related to the model selection is the problem of hypothesis testing for high-dimensional regression. In [ZZ11, Böh12], authors consider null hypotheses of form  $H_{0,i} : \theta_{0,i} = 0$  and propose methods that achieve a given power  $1 - \beta$ , if  $|\theta_{0,i}| > c_\beta\sigma\sqrt{s_0(\log p)/n}$ . Later, [JM14b] proposed a method for random Gaussian designs, with known covariance, under the setting  $s/p \rightarrow \varepsilon$  and  $n/p \rightarrow \delta$ , for constants  $\varepsilon, \delta \in (0, 1)$ . The proposed method achieves a given power  $1 - \beta$ , conditional on that  $|\theta_{0,i}| > c_\beta\sigma/\sqrt{n}$ . The debiasing approach [ZZ14, JM14a, VdGBRD14] also has been proposed to test  $H_{0,i}$  in the high-dimensional setting, with  $s_0 = o(\sqrt{n}/(\log p))$ . In [JM14a], it is shown that the debiasing based framework for testing  $H_{0,i}$  achieves a given power  $1 - \beta$ , if  $\theta_{\min} > c_\beta\sigma\sqrt{(\log p)/n}$ . Applicability of the debiasing approach is extended to the setting of  $s_0 = o(n/(\log p)^2)$ , for random Gaussian designs, using a ‘leave-one-out’ technique [JM18].

#### 1.4. Notations

Here, we provide a summary of notations used throughout this paper. We use  $[p] = \{1, \dots, p\}$  to refer to the first  $p$  integers. For a vector  $v$ , we denote its coordinates by  $v_i$  and let  $v_S$  be the restriction of  $v$  to indices in set  $S$ . Further, the term support of a vector indicates the nonzero coordinates of that vector, i.e.,  $\text{supp}(v) = \{i \in [p] : v_i \neq 0\}$ . We use  $I$  to denote the identity matrix and for clarity we might also make its dimension explicit as in  $I_{d \times d}$ . For a matrix  $A$ , we denote its maximum and minimum singular values by  $\sigma_{\max}(A)$  and  $\sigma_{\min}(A)$ , respectively. For a random vector  $x$ , we denote its subgaussian norm by  $\|x\|_{\psi_2}$  defined as:

$$\|X\|_{\psi_2} \equiv \sup_{q \geq 1} q^{-1/2} (E|X|^q)^{1/q},$$

and for a random vector  $X \in \mathbb{R}^m$ , its subgaussian norm is defined as  $\|X\|_{\psi_2} = \sup_{u \in S^{m-1}} \|\langle X, u \rangle\|_{\psi_2}$ . We use  $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$  to refer to the Gaussian density and  $\Phi(z) = \int_{-\infty}^z \phi(t)dt$  to denote the Gaussian cumulative distribution. For two functions  $f(n)$  and  $g(n)$ , with  $g(n) \geq 0$ , we write  $f(n) = o(g(n))$  if  $g(n)$  grows much faster than  $f(n)$ , i.e.,  $f/g \rightarrow 0$ . We also write  $f(n) = O(g(n))$ , if there exists a positive constant  $C$  such that for all sufficiently large values of  $n$ ,  $|f(n)| \leq C|g(n)|$ .

## 2. FCD Procedure: False Discovery Control via Debiasing

In order to describe FCD framework, we first give an overview of debiasing approach. To this end, we focus on the Lasso estimator [Tib96], given by

$$\hat{\theta}(y, X; \lambda) \equiv \arg \min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\} \quad (9)$$

In case the optimization has more than one optimizer we select one of them arbitrarily. We will often drop the arguments  $y, X$ , as they are clear from the context. There is a vast literature on the properties of the Lasso estimator in the high-dimensional regime ( $n < p$ ), mainly through the lens of point estimation and prediction. A major quantity that plays a key role in the estimation error is the co-called *Compatibility constant* of the design matrix  $X$ . Let  $\hat{\Sigma} \equiv X^\top X/n$  be the sample covariance matrix. In the high-dimensional setting, where  $n < p$ ,  $\hat{\Sigma}$  is always singular, and this makes the estimation of  $\theta_0$  challenging since for the parameter family  $\{\theta = \theta_0 + v\}$ , with  $v$  in the null-space of  $\hat{\Sigma}$ , we have  $X\theta = X\theta_0$  and hence we get the same response vector. A common assumption to cope with this problem is requiring  $\hat{\Sigma}$  to be nonsingular for a restricted set of directions.

**Definition 2.1.** For a symmetric matrix  $\hat{\Sigma} \in \mathbb{R}^{p \times p}$  and a set  $S \subseteq [p]$ , the corresponding compatibility constant is defined as

$$\phi^2(\hat{\Sigma}, S) \equiv \min_{\theta \in \mathbb{R}^p} \left\{ \frac{|S| \langle \theta, \hat{\Sigma} \theta \rangle}{\|\theta_S\|_1^2} : \theta \in \mathbb{R}^p, \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1 \right\}.$$

The matrix  $\hat{\Sigma} \in \mathbb{R}^{p \times p}$  is said to satisfy the *compatibility condition* if  $\phi(\hat{\Sigma}, S) \geq \phi_0$ .

Despite the great properties of Lasso in terms of point estimation and prediction, it is biased due to the  $\ell_1$  penalty term. Indeed, bias is unavoidable in high-dimensional setting ( $n < p$ ) as one needs to produce a point estimate, in  $p$  dimension, from the observed data in lower dimension,  $n$ . Furthermore, characterizing the exact distribution of regularized estimator is in general not tractable. To deal with these challenges, the debiasing approach aims at first removing the bias of Lasso and producing an estimator that is amenable to distributional characterization.

### 2.1. Debiasing Lasso

A debiased estimator  $\hat{\theta}^d$  takes the general simple form of

$$\hat{\theta}^d = \hat{\theta} + \frac{1}{n} M X^\top (y - X\hat{\theta}). \quad (10)$$

Here,  $M$  is a ‘decorrelating’ matrix. There are various proposals for constructing  $M$ ; see e.g. [ZZ14, JM14a, VdGBRD14]. In this paper we use the construction introduced by [JM14a]. Here, we assume that the noise  $w$  is Gaussian and then discuss the non-Gaussian case in Section 2.2.

To set the stage to describe construction of  $M$ , note that by plugging in for  $y = X\theta_0 + w$ , we have

$$\sqrt{n}(\hat{\theta}^{\text{d}} - \theta_0) = \sqrt{n}(M\hat{\Sigma} - \mathbf{I})(\theta_0 - \hat{\theta}) + \frac{1}{\sqrt{n}}MX^{\text{T}}w, \quad (11)$$

where  $\hat{\Sigma} \equiv (X^{\text{T}}X)/n$  is the empirical covariance of the feature vectors. The first term is the bias and is controlled by  $|M\hat{\Sigma} - \mathbf{I}|_{\infty}$ , with  $|\cdot|$  denoting the entrywise  $\ell_{\infty}$  norm. The second term is the unbiased Gaussian noise whose covariance works out at  $M\hat{\Sigma}M^{\text{T}}$ . The decorrelating matrix  $M$  is constructed via a convex optimization that aims at reducing bias and variance of the coordinates of  $\hat{\theta}^{\text{d}}$  at the same time.

Construct  $M = (m_1, m_2, \dots, m_p)^{\text{T}} \in \mathbb{R}^{p \times p}$  by letting  $m_i \in \mathbb{R}^p$  be a solution to the following convex program

$$\begin{aligned} & \text{minimize} && m^{\text{T}}\hat{\Sigma}m, \\ & \text{subject to} && \|\hat{\Sigma}m - e_i\|_{\infty} \leq \mu, \end{aligned} \quad (12)$$

with  $e_i \in \mathbb{R}^p$  being the  $i$ 'th standard unit vector. If any of the above problems is not feasible, we let  $M = \mathbf{I}_{p \times p}$ . Note that  $M$  is constructed solely based on  $X$ . The choice of running parameter  $\mu$  will be discussed in the sequel.

The following proposition proved in [JM14a] shows that the error of the debiased estimator  $\hat{\theta}^{\text{d}}$  can be decomposed as the sum of two ‘bias’ and ‘noise’ terms. In addition, a high probability bound is established on the bias term  $\|\Delta\|_{\infty}$ , which leverages on the properties of the optimization (12) and the estimation error of the Lasso estimator. Note that the compatibility condition for the design matrix  $X$  is required for Lasso to achieve optimal estimation rate in high dimension [BvdG11, vdGB09]. In [JM14a], there is also a version of the following proposition stated for deterministic results, with the compatibility constant  $\phi_0$  explicit in the bound (see Theorem 2.3 therein.) The next proposition concerns the setting of random designs, which per se implies the compatibility condition. Indeed, by employing a reduction principle established by [RZ11], if the population covariance  $\Sigma$  has minimum singular value  $c_{\min} > 0$  and provided a large enough sample size, namely  $n \geq Cs_0 \log(p/s_0)$ , the sample covariance  $\hat{\Sigma}$  satisfies the compatibility condition with constant  $\phi_0 = \sqrt{c_{\min}}/2$ , with high probability.

**Proposition 2.2.** Consider the linear model (2), with gaussian noise,  $w \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_{n \times n})$ , and let  $\hat{\theta}^{\text{d}}$  be the debiased estimator given by Eq. (10), with  $\mu = a\sqrt{(\log p)/n}$ . Then, we have the following decomposition:

$$\sqrt{n}(\hat{\theta}^{\text{d}} - \theta_0) = Z + \Delta, \quad Z|X \sim \mathbf{N}(0, \sigma^2 M\hat{\Sigma}M^{\text{T}}), \quad \Delta = \sqrt{n}(M\hat{\Sigma} - \mathbf{I})(\theta_0 - \hat{\theta}). \quad (13)$$

Consider random design matrices with i.i.d rows and let  $\Sigma = \mathbb{E}(x_1 x_1^{\text{T}})$  be the population level covariance. Suppose that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max}$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . Further, assume that  $X\Sigma^{-1/2}$  has independent subgaussian rows with  $\|\Sigma^{-1/2}x_1\|_{\psi_2} \leq \kappa$ . Then,

choosing  $\lambda = c\sigma\sqrt{(\log p)/n}$ , there exists constant  $C = C(a, \kappa)$ , such that for  $n \geq Cs_0 \log(p/s_0)$ , we have

$$\mathbb{P} \left\{ \|\Delta\|_\infty \geq \left( \frac{16ac_0\sigma}{c_{\min}} \right) \frac{s_0 \log p}{\sqrt{n}} \right\} \leq 4e^{-c_1 n} + 4p^{-c_2}, \quad (14)$$

where  $c_1$  and  $c_2$  are constants depending on  $\kappa, a, c_0, c_{\min}, c_{\max}$ .

The next lemma controls the variance of the noise coordinates  $Z_i$  in terms of the diagonal entries of the precision matrix.

**Lemma 2.3** ([JM14a]). Let  $\Omega \equiv \Sigma^{-1}$  be the precision matrix. Under the assumption of Proposition 2.2, the following holds true for any fixed sequence of integers  $i = i(n)$ :

$$\mathbb{P} \left( m_i^\top \widehat{\Sigma} m_i - \Omega_{i,i} \geq \epsilon \right) \leq 2e^{-(n/6)(\epsilon/\epsilon\kappa')^2} + 2p^{-c}, \quad (15)$$

for  $\kappa' \equiv 2\kappa^2 c_{\min}^{-1}$  and a constant  $c = c(a) > 0$ .

## 2.2. Extension to Non-Gaussian Noise

In the decomposition (2.2), we have  $Z = MX^\top W/\sqrt{n}$  and given that  $W \sim \mathbf{N}(0, \Sigma)$ , we have  $Z|X \sim \mathbf{N}(0, \sigma^2 M \widehat{\Sigma} M^\top)$ . In [JM14a], it is shown that by a slight modification of optimization (12),  $Z$  admits the same conditional distribution even for non-Gaussian noise. For the reader's convenience and to be self-contained we briefly explain it here.

Note that for any fixed  $i \in [p]$ , we have

$$Z_i = \frac{1}{\sqrt{n}} \sum_{j=1}^n \xi_j, \quad \text{with} \quad \xi_j \equiv \frac{m_i^\top x_j w_j}{\sigma(m_i^\top \widehat{\Sigma} m_i)^{1/2}}.$$

Conditional on  $X$ , the terms  $\xi_j$  are zero mean and independent. Moreover,  $\sum_{j=1}^n \mathbb{E}(\xi_j^2|X) = n$ . Therefore, if the Lindeberg's condition holds, that is to say for every  $\epsilon > 0$ , almost surely

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \mathbb{E}(\xi_j^2 \mathbb{I}(|\xi_j| > \epsilon\sqrt{n})|X) = 0,$$

then  $\sum_{j=1}^n \xi_j/\sqrt{n}|X \xrightarrow{d} \mathbf{N}(0, 1)$ . The construction of  $M$  can slightly be modified to ensure the Lindeberg's condition, namely optimization problem (12) should be modified as follows:

$$\begin{aligned} & \text{minimize} && m^\top \widehat{\Sigma} m, \\ & \text{subject to} && \|\widehat{\Sigma} m - e_i\|_\infty \leq \mu, \\ & && \|Xm\|_\infty \leq n^\beta \quad \text{for arbitrary fixed } 0 < \beta < 1/2 - a^{-1}, \end{aligned} \quad (16)$$

where we recall the parameter  $a$  from Eq. (3). The following lemma, which is from [JM14a], shows that by this modification, the marginals of  $Z_j$  are asymptotically normal.

**Lemma 2.4** ([JM14a], Theorem 4.1). Under conditions (3) on the noise term, and using optimization (12) to construct  $M$ , we have that  $Z_i|X \xrightarrow{d} N(0, 1)$ .

The above result can be easily generalized to fixed-dimensional marginals of  $Z$ , by using the fact that a vector has a multivariate normal distribution if every linear combination of its coordinates is normally distributed.

With this overview of debiasing approach we are ready to explain the FCD procedure.

### 2.3. FCD Procedure

#### 2.3.1. Construction of Test Statistics

In order to construct the test statistics, we first need to propose a consistent estimate of noise variance,  $\sigma^2$ . There are already several proposals for this in the literature. See e.g., [FL01, FL08, SBvdG10, Zha10, SZ12, BC13, FGH12, RTF16]. To be concrete, we use the scaled Lasso [SZ12] given by

$$\{\hat{\theta}, \hat{\sigma}\} \equiv \arg \min_{\theta \in \mathbb{R}^p, \sigma > 0} \left\{ \frac{1}{2\sigma n} \|y - X\theta\|_2^2 + \frac{\sigma}{2} + \bar{\lambda} \|\theta\|_1 \right\} \quad (17)$$

We state the following lemma that shows  $\hat{\sigma}$  is a consistent estimate of  $\sigma$ . We refer to [JM14a, Lemma 3.3] or [SZ12, Theorem 1] for its proof.

**Lemma 2.5.** Consider a sequence of design matrices  $X \in \mathbb{R}^{n \times p}$ , with dimensions  $n \rightarrow \infty$ ,  $p = p(n) \rightarrow \infty$ . For each  $n$ , let  $\Sigma \in \mathbb{R}^{p \times p}$  such that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max} < \infty$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . Further, assume that  $X\Sigma^{-1/2}$  has independent subgaussian rows, with zero mean and subgaussian norm  $\|\Sigma^{-1/2}x_1\|_{\psi_2} \leq \kappa$ . Let  $\hat{\sigma}$  be the scaled Lasso estimate of the noise level, defined by (17), with  $\bar{\lambda} = 2\sqrt{(2 \log p)/n}$ . Then, assuming  $s_0 = o(n/(\log p))$ , the estimator  $\hat{\sigma}$  satisfies the following relation:

$$\lim_{n \rightarrow \infty} \sup_{\theta_0 \in \mathbb{R}^p, \|\theta_0\|_0 \leq s_0} \mathbb{P} \left( \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon \right) = 0.$$

Here,  $\mathbb{P}$  is w.r.t the randomness of the noise  $w$  and the design  $X$ .

Define  $\Lambda = M\hat{\Sigma}M^\top$ . For  $i \in [p]$ , we define test statistic  $T_i$  as follows:

$$T_i \equiv \frac{\sqrt{n}\hat{\theta}_i^d}{\hat{\sigma}\sqrt{\Lambda_{ii}}}. \quad (18)$$

For a given threshold level  $t \geq 0$ , we reject  $H_{0,i}$  if  $|T_i| \geq t$  and we return sign of  $T_i$  as the estimate of sign of  $\theta_{0,i}$ . We also let  $R(t) = \sum_{i=1}^p \mathbb{I}(|T_i| \geq t)$  be the total set of rejections at threshold  $t$ . Next, we discuss how to choose a data dependent threshold  $t$  to ensure that directional FDR and FDP are controlled at a pre-assigned level  $q \in [0, 1]$ .

### 2.3.2. A Data Dependent Threshold for the Test Statistics

- Step 1: For the pre-assigned level  $q \in [0, 1]$ , let  $t_p = (2 \log p - 2 \log \log p)^{1/2}$  and calculate

$$t_0 = \inf \left\{ 0 \leq t \leq t_p : \frac{2p(1 - \Phi(t))}{R(t) \vee 1} \leq q \right\}. \quad (19)$$

If (19) does not exist then set  $t_0 = \sqrt{2 \log p}$ .

- Step 2: For  $i \in [p]$ , reject  $H_{0,i}$  if  $|T_i| \geq t_0$ .
- Step 3: We return  $\widehat{\text{sign}}_i = \text{sign}(T_i)$  as the estimate of  $\text{sign}(\theta_{0,i})$ .

## 3. Main Results

### 3.1. Control of Directional False Discovery Rate

Suppose that the design matrix  $X$  has i.i.d rows with  $\Sigma = \mathbb{E}(x_1 x_1^\top)$  being the population covariance. Let  $\Omega \equiv \Sigma^{-1}$  be the precision matrix and recall the definition  $\Lambda \equiv M \widehat{\Sigma} M^\top$ , where  $M$  is the decorrelating matrix used in construction of the debiased estimator.

We also define the normalized matrices  $\Omega^0$  and  $\Lambda^0$  as

$$\Omega_{ij}^0 = \frac{\Omega_{ij}}{\sqrt{\Omega_{ii} \Omega_{jj}}}, \quad \Lambda_{ij}^0 = \frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii} \Lambda_{jj}}}. \quad (20)$$

For a given constant  $\gamma > 0$ , define

$$\Gamma(\gamma, c_0) \equiv \left\{ (i, j) : 1 \leq i, j \leq p, |\Omega_{ij}^0| \geq c_0 (\log p)^{-2-\gamma} \right\}, \quad (21)$$

for some constant  $c > 0$ . The following theorem states a guarantee on the directional false discovery rate of the FCD procedure introduced in the previous section.

**Theorem 3.1.** Consider random design matrices with i.i.d rows and let  $\Sigma = \mathbb{E}(x_1 x_1^\top)$  be the population level covariance. Suppose that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max}$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . In addition, assume that  $X \Sigma^{-1/2}$  has independent subgaussian rows with  $\|\Sigma^{-1/2} x_1\|_{\psi_2} = \kappa$ . Also assume that:

- (i)  $s_0 = o(\sqrt{n}/(\log p)^2)$ .
- (ii) There exist positive constants  $c_0, \gamma$ , such that  $|\Gamma(\gamma, c_0)| = o(p^{1+\rho})$ , for some constant  $\rho \in [0, 1)$ .
- (iii) We have  $|\{(i, j) : |\Omega_{ij}^0| > (1 - \rho)/(1 + \rho)\}| = O(p)$ .

Then, for FCD procedure we get

$$\limsup_{(n,p) \rightarrow \infty} \text{FDR}_{\text{dir}} \leq q. \quad (22)$$

Further, for any  $\varepsilon > 0$ ,

$$\lim_{(n,p) \rightarrow \infty} \mathbb{P}(\text{FDP}_{\text{dir}} \leq q + \varepsilon) = 1. \quad (23)$$

**Remark 3.2.** While directional FDR is the *expected* directional false discovery proportion ( $\text{FDP}_{\text{dir}}$ ), it is idealized for a variable selection procedure to control  $\text{FDP}_{\text{dir}}$  in any given realization. In general, controlling  $\text{FDR}_{\text{dir}}$  does not control the variations of  $\text{FDP}_{\text{dir}}$ . As noted by [Owe05], the variance of FDP can be large if the test statistics are correlated, which is the case here. Let us emphasize that by Eq. (23), our FCD controls  $\text{FDP}_{\text{dir}}$ , with high probability.

**Examples.** Here, we provide several examples of the precision matrices that satisfy conditions (ii)-(iii) of Theorem 3.1 to demonstrate its applicability.

*Example 1:* Our first example is the circulant covariance matrices, where  $\Sigma_{ij} = \eta^{|i-j|}$ , for some constant  $\eta \in (0, 1)$ . It is simple to see that the inverse of such matrices has at most three nonzero coordinates per row. Therefore, the conditions will be satisfied by choosing  $\rho = 1$ , and  $c > 0, \gamma < \infty$ , arbitrary.

*Example 2:* Suppose that  $\Sigma$  is block diagonal with size of blocks to bounded (as  $p \rightarrow \infty$ ). Then, the precision matrix will also have a block diagonal structure with blocks of bounded size. It is easy to check conditions, with choosing  $\rho = 1$  and  $c > 0, \gamma < \infty$ , arbitrary.

*Example 3:* Consider the equi-correlated features, where  $\Sigma = (1 - r)\mathbf{I} + r\mathbf{1}\mathbf{1}^\top$ , for some constant  $r \in (0, 1)$ , where  $\mathbf{1} \in \mathbb{R}^p$  denotes the all-one vector. Then, we have  $\Omega = (a - b)\mathbf{I} + b\mathbf{1}\mathbf{1}^\top$ , with

$$a = \frac{(p-2)r+1}{(p-2)r - (p-1)r^2 + 1}, \quad b = \frac{-r}{(p-2)r - (p-1)r^2 + 1}. \quad (24)$$

Note that  $|b| = O(1/p)$ . Therefore, the conditions hold for arbitrary constants  $c > 0, 0 < \rho < 1$ .

Finally, consider two matrices  $\Omega^{(1)}$  and  $\Omega^{(2)}$ , with same diagonal entries  $\Omega_{ii}^{(1)} = \Omega_{ii}^{(2)}$ , for  $i \in [p]$ , such that  $\Omega^{(1)}$  dominates  $\Omega^{(2)}$  on off-diagonal entries, i.e.,  $\Omega_{ij}^{(1)} \geq \Omega_{ij}^{(2)}$ , for  $i \neq j \in [p]$ . Then it is easy to see that if  $\Omega^{(1)}$  satisfies Conditions (i)-(ii), so does  $\Omega^{(2)}$ .

### 3.2. Power Analysis

Recall that  $S_0 \equiv \text{supp}(\theta_0)$  is the set of indices of the truly significant features. Let  $\widehat{S}$  denote the set of significant parameters returned by our FCD procedure, namely

$$\widehat{S} = \{1 \leq j \leq p : |T_j| \geq t_0\}. \quad (25)$$

The power of a selected model  $\widehat{S}$  is defined as

$$\text{Power}(\widehat{S}) = \mathbb{E} \left[ \frac{|\{j \in \widehat{S} : \widehat{\text{sign}}_j = \text{sign}(\theta_{0,j})\}|}{\max(|\widehat{S}|, 1)} \right]. \quad (26)$$

We are now ready to characterize the statistical power of the FCD procedure for the linear model (2).

**Theorem 3.3.** Consider a sequence of random design matrices  $X \in \mathbb{R}^{n \times p}$ , with dimension  $n \rightarrow \infty$ ,  $p = p(n) \rightarrow \infty$  and  $\Sigma = \mathbb{E}(x_1 x_1^\top) \in \mathbb{R}^{p \times p}$ . Suppose that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max}$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . Further, assume that  $X \Sigma^{-1/2}$  has independent subgaussian rows with  $\|\Sigma^{-1/2} x_1\|_{\psi_2} = \kappa$ . Suppose that  $s_0 = o(\sqrt{n}/(\log p)^2)$  and for  $i \in S = \text{supp}(\theta_0)$ , we have  $|\theta_{0,i}| > (\sigma/\sqrt{n})\sqrt{2\Omega_{ii} \log(p/s_0)}$ . Then, the following holds true:

$$\liminf_{n \rightarrow \infty} \frac{\text{Power}(\widehat{S})}{1 - \beta(\theta_0, n)} \geq 1 \quad (27)$$

$$1 - \beta(\theta_0, n) = \frac{1}{s_0} \sum_{i \in S} F \left( \frac{q s_0}{p}, \frac{\sqrt{n} |\theta_{0,i}|}{\sigma \sqrt{\Omega_{ii}}} \right), \quad (28)$$

where, for  $\alpha \in [0, 1]$  and  $u \in \mathbb{R}_+$ , the function  $F(\alpha, u)$  is defined as follows:

$$F(\alpha, u) \equiv 1 - \Phi(\Phi^{-1}(1 - \alpha/2) - u). \quad (29)$$

We refer to Section 7.2 for the proof of Theorem 3.3.

**Corollary 3.4.** It is easy to see that for any fixed  $\alpha \in [0, 1]$ , function  $u \mapsto F(\alpha, u)$  is monotone increasing. Therefore, as a result of Theorem 3.3, we have

$$\liminf_{n \rightarrow \infty} \frac{\text{Power}(\widehat{S})}{F \left( \frac{q s_0}{p}, \frac{\sqrt{n} \theta_{\min}}{\sigma \sqrt{\Omega_{ii}}} \right)} \geq 1. \quad (30)$$

**Corollary 3.5.** Under the assumptions of Theorem 3.3, if

$$\sqrt{n} \theta_{\min} - \sigma \sqrt{2 \max_{i \in [p]} (\Omega_{ii} \log(2p/(q s_0)))} \rightarrow \infty,$$

then  $\text{Power}(\widehat{S}) \rightarrow 1$ , as  $n \rightarrow \infty$ .

Proof of Corollary 3.5 is given in Appendix A.4.

#### 4. Improved Results for Gaussian Designs

In [JM18], the authors improved upon Proposition 2.2 for Gaussian designs by providing a sharper bound for  $\|\Delta\|_\infty$  using a ‘leave-one-out’ technique. Specifically, for Gaussian designs with known population covariance, it is shown that  $\|\Delta\|_\infty = o_p(\sqrt{\frac{s_0}{n}} \log p)$ . The same bound holds when the population covariance is unknown but can be estimated sufficiently well. e.g., if the inverse covariance is sufficiently sparse. In this section, we aim at employing this result to relax the sparsity assumption (Condition (i)) in Theorem 3.1.

### 4.1. Known Covariance

Consider linear model (2) where the design  $X$  has independent Gaussian rows, with zero mean and covariance  $\Sigma$ . Also, denote by  $\Omega \equiv \Sigma^{-1}$  be the inverse population covariance, a.k.a precision matrix. Here, we assume that  $\Sigma$  is known and consider the test statistic  $T_i$ , given by (18) where  $\hat{\theta}^d$  is the debiased estimator with  $M = \Omega$ .

For an integer  $1 \leq k \leq p$ , define  $\tau(\Omega, k)$  as follows:<sup>2</sup>

$$\tau(\Sigma, k) \equiv \max_{A \subseteq [p], |A| \leq k} \|(\Sigma_{A,A})^{-1}\|_\infty,$$

where  $\|\cdot\|_\infty$  denotes the  $\ell_\infty$  operator norm (maximum  $\ell_1$  norm of the rows). As proved in [JM18], we have the following bound in place:

$$\tau(\Sigma, k) \leq \min \left\{ \|\Omega\|_\infty, \sqrt{k} \sigma_{\max}(\Omega) \right\}.$$

The next theorem is analogues to Theorem 3.1 for Gaussian designs, under a weaker assumption on the sparsity level  $s_0$ .

**Theorem 4.1.** (Known covariance). Consider a sequence of Gaussian random design matrices  $X \in \mathbb{R}^{n \times p}$ , with dimension  $n \rightarrow \infty$ ,  $p = p(n) \rightarrow \infty$ . Suppose that  $X$  has i.i.d rows with zero mean and  $\Sigma = \mathbb{E}(x_1 x_1^\top)$  be the population covariance. Suppose that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max}$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . Further, assume that:

- (i)  $s_0 = o(n/(\log p)^4)$ .
- (ii) Let  $C_0 = (32c_{\max}/c_{\min}) + 1$ . We have  $\tau(\Sigma, C_0 s_0) \leq \tau_0$ , for some constant  $\tau_0 > 0$ .
- (iii) There exist positive constants  $c_0, \gamma$ , such that  $|\Gamma(\gamma, c_0)| = o(p^{1+\rho})$ , for some constant  $\rho \in [0, 1)$ .
- (iv) We have  $|\{(i, j) : |\Omega_{ij}^0| > (1 - \rho)/(1 + \rho)\}| = O(p)$ .

Then, for FCD procedure we get

$$\limsup_{(n,p) \rightarrow \infty} \text{FDR}_{\text{dir}} \leq q. \tag{31}$$

Further, for any  $\varepsilon > 0$ ,

$$\lim_{(n,p) \rightarrow \infty} \mathbb{P}(\text{FDP}_{\text{dir}} \leq q + \varepsilon) = 1. \tag{32}$$

The proof of Theorem 4.1 proceeds along the same lines as proof of Theorem 3.1 and uses the result of [JM18, Theorem 3.8]. We refer to section 7.3 for its proof.

---

<sup>2</sup>In [JM18], the authors use the notation  $\rho(\Omega, k)$  to refer to the same quantity. We avoid that notation as we have used the symbol  $\rho$  in Condition (iii) in Theorem 4.1.

### 4.2. Unknown Covariance

For the case of unknown covariance, we follow the construction of the decorrelating matrix  $M$  proposed in [VdGBRD14]. This construction is based on node-wise Lasso on matrix  $X$ . Formally, for  $i \in [p]$ , let  $\tilde{x}_i$  be the  $i$ -th column of  $X$  and represent it via sparse regression against all other columns:

$$\hat{\gamma}_i(\tilde{\lambda}) = \arg \min_{\gamma \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\tilde{x}_i - X_{\sim i} \gamma\|_2^2 + \tilde{\lambda} \|\gamma\|_1 \right\},$$

where  $X_{\sim i}$  is the submatrix obtained by removing the  $i$ -th column. Let

$$\hat{C} = \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{bmatrix}.$$

Also define

$$\hat{T}^2 = \text{diag}(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2), \quad \hat{\tau}_i^2 = \frac{1}{n} (\tilde{x}_i - X_{\sim i} \hat{\gamma}_i)^\top \tilde{x}_i. \quad (33)$$

The decorrelating matrix  $M$  is then defined as

$$M \equiv \hat{T}^{-2} \hat{C}. \quad (34)$$

We consider the FDC procedure, where the test statistic  $T_i$  is given by (18) and  $\hat{\theta}^d$  is the debiased estimator with the decorrelating matrix  $M$  (34).

Define the sparsity level  $s_\Omega$  for the precision matrix  $\Omega$  as:

$$s_\Omega \equiv \max_{i \in [p]} |\{j \neq i, \Omega_{i,j} \neq 0\}|.$$

In words,  $s_\Omega$  is the maximum sparsity of the rows of  $\Omega$ .

For the case of Gaussian designs with unknown covariance, we prove that the directional FDR of the FCD procedure is controlled under a weaker assumption on the sparsity of the parameters  $s_0$ , provided  $s_\Omega$  is small enough.

**Theorem 4.2.** (Unknown covariance). Consider a sequence of Gaussian random design matrices  $X \in \mathbb{R}^{n \times p}$ , with dimension  $n \rightarrow \infty$ ,  $p = p(n) \rightarrow \infty$  and  $X$  has i.i.d rows with zero mean and covariance  $\Sigma$ . Assume that  $\sigma_{\min}(\Sigma) \geq c_{\min} > 0$  and  $\sigma_{\max}(\Sigma) < c_{\max}$ , for some constants  $c_{\min}$ ,  $c_{\max}$  and  $\max_{i \in [p]} \Sigma_{ii} \leq 1$ . Further, suppose that

$$(i) \quad s_0 = o(n/(\log p)^4) \text{ and } \min(s_0, s_\Omega) = o(\sqrt{n}/(\log p)^2).$$

and Conditions (ii), (iii), (iv) in Theorem 4.1 hold for  $\Sigma$ . Then, for FCD procedure we get

$$\limsup_{(n,p) \rightarrow \infty} \text{FDR}_{\text{dir}} \leq q. \quad (35)$$

Further, for any  $\varepsilon > 0$ ,

$$\lim_{(n,p) \rightarrow \infty} \mathbb{P}\left(\text{FDP}_{\text{dir}} \leq q + \varepsilon\right) = 1. \quad (36)$$

The proof of Theorem 4.2 proceeds along the same lines as proof of Theorem 3.1 and uses the result of [JM18, Theorem 3.13]. We refer to section 7.4 for its proof.

## 5. Numerical Experiments

We consider linear model (2) where the design matrix  $X$  is generated by drawing its rows independently from  $\mathbf{N}(0, \Sigma)$ . The covariance  $\Sigma \in \mathbb{R}^{p \times p}$  has a circulant structure with  $\Sigma_{ij} = \eta^{|i-j|}$ , for some constant  $\eta \in (0, 1)$ . We then normalize the columns of  $X$  to have unit norm. We generate the vector of coefficients  $\theta_0 \in \mathbb{R}^p$  by choosing a subset of indices  $S \subseteq [p]$  at random, of size  $s_0$  and setting  $\theta_{0,i}$  from  $\{\pm A\}$  uniformly at random and  $\theta_{0,i} = 0$ , for  $i \notin S_0$ . The noise term  $W$  is drawn from  $\mathbf{N}(0, \mathbf{I}_{n \times n})$ .

We perform three sets of simulations to compare the performance of FCD procedure with model free knockoff and to examine the effects of sparsity level, signal magnitude, and feature correlation. We also compare the empirical power of FCD with the analytical lower bound provided in Corollary 3.4. In all simulations, we set the target level FDR to  $q = 0.1$ .

For FCD procedure, we use the implementation of the debiased method provided by <http://web.stanford.edu/montanar/sslasso/>, to construct the debiased estimator. For model free knockoff, we use the implantation provided by <http://web.stanford.edu/group/candes/knockoffs/>.

**Effect of Signal Amplitude:** We choose  $n = 2000$ ,  $p = 3000$ ,  $k = 100$ ,  $\eta = 0.1$  and vary the signal amplitude in the set  $A \in \{0.5, 1, 1.5, \dots, 5.5, 6\}$ . For the FCD procedure and the model free knockoff, we compute the directional FDR and power by averaging across 100 realizations of noise and the generation of coefficient vector  $\theta_0$ . The results are plotted in Figure 1. As we observe, both methods control  $\text{FDR}_{\text{dir}}$  under the target level  $q = 0.1$ . As expected, the power of both procedures increases as the signal amplitude increases, with FCD procedure having larger power than the knockoff method over the entire range of signal amplitudes. The FCD procedure turn out to be more powerful than knockoff procedure.

We also plot the analytical lower bound on the power of FCD procedure, given in Corollary 3.4. As we see the lower bound is quite close to the actual empirical power of FCD procedure in the setup tested.

**Effect of feature correlation:** We test the effect of feature correlations on the performance of FCD procedure, comparing it with the model free knockoff. We set  $n = 700$ ,  $p = 1000$ ,  $k = 50$ ,  $A = 4.5$ . Recall that the rows of the design

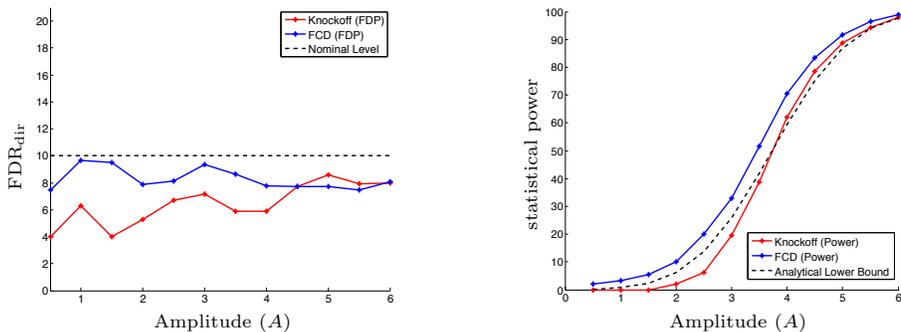


FIG 1. Testing FCD and model free knockoff methods with varying the coefficients amplitude  $A$ . Here,  $n = 2000$ ,  $p = 3000$ ,  $k = 100$ ,  $\eta = 0.1$ . The target level is  $q = 10\%$ .  $FDR_{dir}$  and power are computed by averaging over 100 realizations of noise and coefficient vectors.

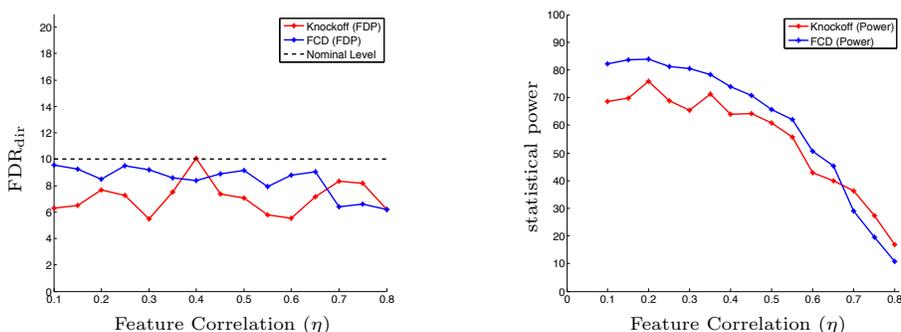


FIG 2. Testing FCD and model free knockoff methods with varying the feature correlation parameter  $\eta$ . Here,  $n = 700$ ,  $p = 1000$ ,  $k = 50$ ,  $A = 4.5$ . The target level is  $q = 10\%$ .  $FDR_{dir}$  and power are computed by averaging over 100 realizations of noise and design matrices.

matrix  $X$  are generated from a  $N(0, \Sigma)$  distribution, with  $\Sigma_{ij} = \eta^{|i-j|}$ , and then the columns of  $X$  are normalized to have unit norm. We vary the parameter  $\eta$  in the set  $\{0.1, 0.15, 0.2, \dots, 0.75, 0.8\}$ . For each value of  $\eta$ , we compute  $FDR_{dir}$  and power for both methods, by averaging over 100 realizations of noise and design matrix  $X$ . The results are displayed in Figure 2.

As observed, both methods control  $FDR_{dir}$  over the range of correlations tested. From the power plot, we see that the power of both methods decays as the features correlations increase. This is expected because when the features are highly correlated it is harder to distinguish between them and report the truly significant ones. Indeed, for large values of  $\eta$ , both methods select a few variables. This way,  $FDR_{dir}$  is still controlled but the power is low. The proposed FCD procedure has higher power than model free knockoff for  $\eta \leq 0.65$ .

**Effect of Sparsity:** Here, we set  $n = 2000$ ,  $p = 3000$ ,  $A = 4.5$ ,  $\eta = 0.1$  and vary the sparsity level of the coefficients in the set  $k \in \{10, 15, 20, \dots, 130\}$ . For

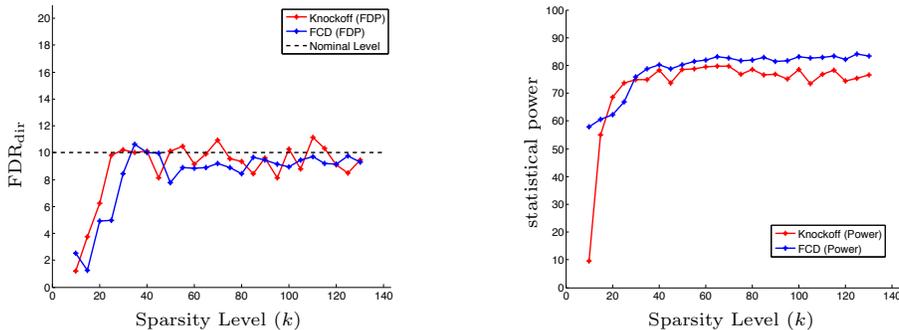


FIG 3. Testing FCD and model free knockoff methods with varying the sparsity level  $k$ . Here,  $n = 2000$ ,  $p = 3000$ ,  $A = 4.5$ ,  $\eta = 0.1$ . The target level is  $q = 10\%$ .  $\text{FDR}_{\text{dir}}$  and power are computed by averaging over 100 realizations of noise and coefficient vectors.

both methods, the power and FDR are computed by averaging over 100 trials of noise and the generation of coefficient vector  $\theta_0$ . Both methods control  $\text{FDR}_{\text{dir}}$  over the entire range, with FCD achieving lower  $\text{FDR}_{\text{dir}}$  for small values of  $k$ . In terms of power, both methods have close power, and the FCD has higher power for small  $k$ .

## 6. Real Data Experiments

In this section we evaluate the proposed method to find the mutations in the Human Immunodeficiency Virus Type 1 associated with drug resistance<sup>3</sup>. This dataset is presented and analyzed in [RTW<sup>+</sup>06] and is obtained by analyzing HIV-1 subtype B sequences from persons with histories of antiretroviral treatment. The dataset contains the mutations in the protease and reverse transcriptase (RT) positions of the HIV-1 subtype B sequences which correspond to resistance to Protease Inhibitors (PI), to nucleoside reverse transcriptase inhibitors (NRTIs) and to non-nucleoside RT inhibitors (NNRTIs).

In order to deal with missing measurements and preprocessing the dataset we mostly follow the steps taken in [BC15]. The design matrix  $X \in \{0, 1\}^{n \times p}$  is formed by letting  $X_{ij} = 1$  if the  $i$ 'th sample contains the  $j$ 'th mutation and  $X_{ij} = 0$  otherwise. Further, for a specific drug, the  $i$ 'th entry of the response vector  $y_i$  denotes the logarithm of the increase in the resistance to that drug in the  $i$ 'th patient. We let  $q = 0.2$  and we apply the FCD procedure described in subsection 2.3 to detect the mutations in the HIV-1 associated with resistance to each drug. In order to evaluate the performance of our method, we compare it with the knockoff filter procedure [BC15] with the test statistics based on lasso. The size of the dataset ( $n$ ,  $p$ ) for each drug is noted under the bar plot corresponding to that drug. For all cases, except the data for resistance to TDF, we have  $n > 2p$ .

<sup>3</sup>The dataset is available online at [https://hivdb.stanford.edu/pages/published\\_analysis/genophenoPNAS2006](https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006).

We have used two different methods for generating the knockoff variables; in knockoff1, the knockoff variables are generated by solving a semi-definite program (SDP) and in knockoff2, equi-correlated knockoff variables are created without solving an SDP at a lower computational cost<sup>4</sup>. Since this is a real data experiment, there is no ground truth. However, we use the methodology in [BC15] to assess our results. In order to do this, we evaluate the reproducibility of the outcomes of these procedures by comparing them with treatment-selected mutation (TSM) panels provided in [RFZ<sup>+</sup>05]. These panels contain mutations that are observed more frequently in virus samples from patients that have been treated by each drug in compare with the patients who have never been treated with that drug. Since these panels are created independently from the dataset that we use, they can provide a good measure for validating the reproducibility of the results obtained by each procedure.

A summary of the results can be seen in Figures 4, 5, 6. It can be seen that the FCD method achieves the target FDR level of  $q = 0.2$  and the obtained power in half of the cases (8 out of 16 drugs) is larger than the power achieved by the knockoff filter. Overall, the achieved power is comparable with the power of the knockoff filter method.

## 7. Proof of Main Theorems

### 7.1. Proof of Theorem 3.1

Define  $G(t) = 2(1 - \Phi(t))$ , with  $\Phi(t)$  denoting the standard Gaussian cumulative distribution. We start by two lemmas about the properties of  $G(t)$ .

**Lemma 7.1.** For all  $t \geq 0$ , we have

$$\frac{2}{t + 1/t} \phi(t) < G(t) < \frac{2}{t} \phi(t), \quad (37)$$

where  $\phi(t) = e^{-t^2/2}/\sqrt{2\pi}$  is the standard Gaussian density.

Lemma 7.1 is the standard tail bound on the Gaussian distribution and its proof is omitted.

**Lemma 7.2.** For all  $t > 0$ ,  $\varepsilon < \min(1, 1/t)$  and  $\delta < \min(1, 1/t^2)$ , the following holds true:

$$\frac{G((1 - \delta)t - \varepsilon)}{G(t)} \leq 1 + 8(\varepsilon + \varepsilon t + \delta + \delta t^2). \quad (38)$$

Proof of Lemma 7.2 is given in Appendix A.1.

Using Proposition (2.2), we have

$$T_i = \frac{\sqrt{n}\theta_{0,i}}{\hat{\sigma}\sqrt{\Lambda_{ii}}} + \frac{\sigma}{\hat{\sigma}} \tilde{Z}_i + \frac{\Delta_i}{\hat{\sigma}\sqrt{\Lambda_{ii}}} \quad (39)$$

<sup>4</sup>More information regarding the procedure are available at <https://web.stanford.edu/groupcandes/knockoffs/software/knockoff/index.html>.

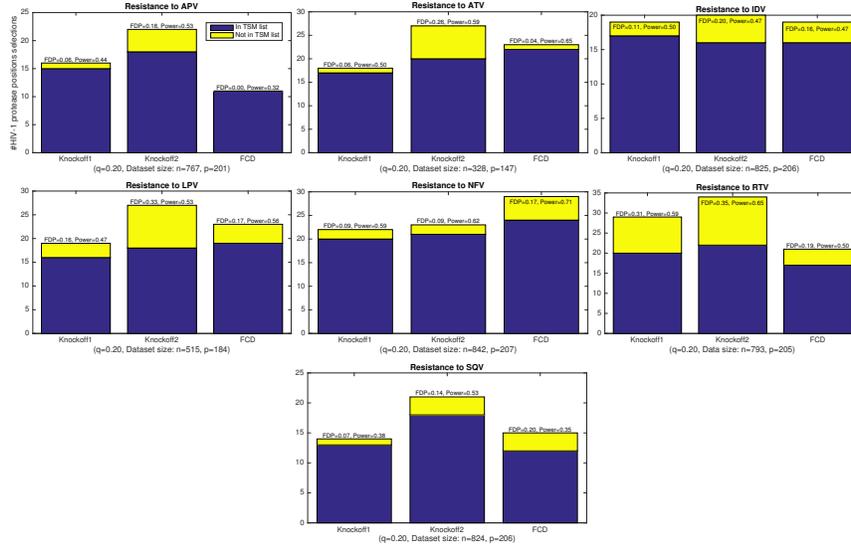


FIG 4. Summary of the results of applying the knockoff filter and FCD for detecting the mutation positions in HIV-1 associated with resistance to type-PI drugs using the dataset in [RTW<sup>+</sup>06]. In these experiments we have used  $q = 0.2$ . In the plots, blue bars show the number of detected positions by different methods that appear in the TSM panels. On top of each bar the proportion of detected mutations that appear in the TSM panel (an estimate for FDP) and the proportion of mutations in the TSM panel that are detected by different methods (an estimate for power) are stated.

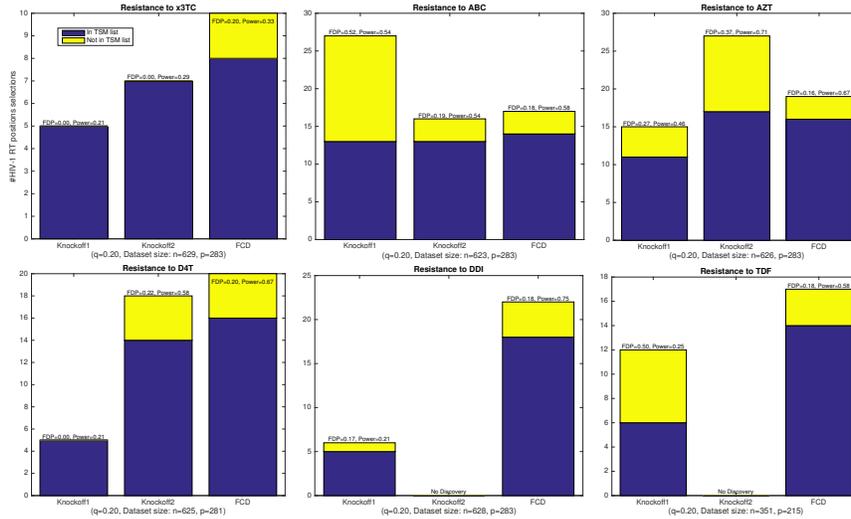


FIG 5. Same as Figure 4 for type-NRTI drugs.

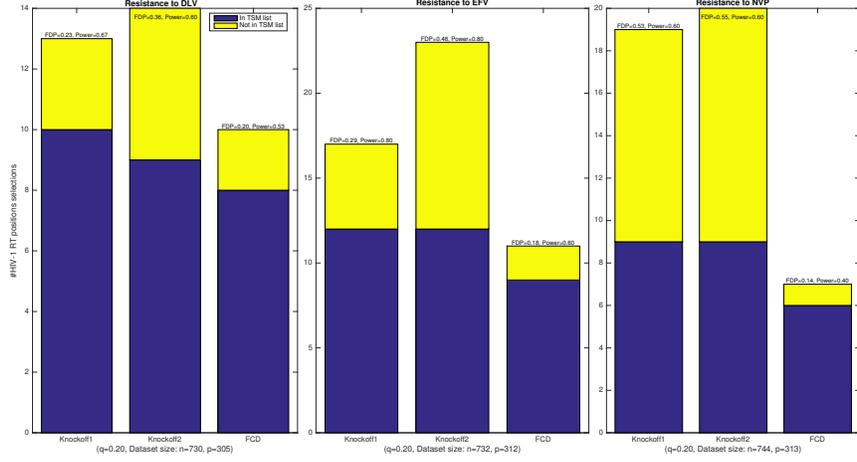


FIG 6. Same as Figure 4 for type-NNRTI drugs.

where  $\tilde{Z}_i \sim N(0, \Lambda^0)$ . By invoking [JM14a, Lemma 3.1],  $\Lambda_{ii}$  are bounded from below by an arbitrary fixed constant  $0 < c < 1$ , for large enough  $n$ . In addition, since  $|\Lambda^0 - \Omega^0|_\infty = o_p(1)$ , for  $(i, j) \in \Gamma(\gamma, c_0)^c$  we have

$$|\Lambda_{ij}^0| < C(\log p)^{-2-\gamma}, \quad (40)$$

for some constant  $C > 0$ . Further, by Condition (iii) in the theorem statement, we have

$$\left| \left\{ (i, j) : |\Omega_{ij}^0| > \frac{1-\rho}{1+\rho} \right\} \right| = O(p). \quad (41)$$

Define  $S_{\geq 0} \equiv \{i \in [p] : \theta_{0,i} \geq 0\}$  and  $S_{\leq 0} \equiv \{i \in [p] : \theta_{0,i} \leq 0\}$ .

We first consider the case that  $t_0$ , given by (19), does not exist. In this case,  $t_0 = \sqrt{2 \log p}$  and for any  $\varepsilon > 0$  we have

$$\mathbb{P} \left( \sum_i \mathbb{I}(\widehat{\text{sign}}_i \neq \text{sign}(\theta_{0,i})) \geq 1 \right) \leq \mathbb{P} \left( \sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq \sqrt{2 \log p}) \geq 1 \right) \quad (42)$$

$$+ \mathbb{P} \left( \sum_{i \in S_{\geq 0}} \mathbb{I}(T_i \leq -\sqrt{2 \log p}) \geq 1 \right). \quad (43)$$

We can bound the first term on the right hand side above as

$$\begin{aligned}
& \mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq \sqrt{2 \log p}) \geq 1\right) \\
& \leq \mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}\left(\frac{\sigma}{\hat{\sigma}} \tilde{Z}_i + \frac{\Delta_i}{\hat{\sigma} \sqrt{\Lambda_{ii}}} \geq \sqrt{2 \log p}\right) \geq 1\right) \\
& \leq \mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}\left(\tilde{Z}_i \geq \frac{\hat{\sigma}}{\sigma} \sqrt{2 \log p} - \frac{\|\Delta\|_{\infty}}{\sigma \sqrt{c}}\right) \geq 1\right) \\
& \leq p \max_{i \in [p]} \mathbb{P}\left(\tilde{Z}_i \geq (1 - \varepsilon) \sqrt{2 \log p} - \varepsilon\right) \\
& \quad + \mathbb{P}\{\|\Delta\|_{\infty} \geq \sigma \varepsilon \sqrt{c}\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon\right\} \\
& \leq \frac{p}{2} G\left((1 - \varepsilon) \sqrt{2 \log p} - \varepsilon\right) \\
& \quad + \mathbb{P}\{\|\Delta\|_{\infty} \geq \sigma \varepsilon \sqrt{c}\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon\right\}.
\end{aligned}$$

which goes to zero as  $n, p \rightarrow \infty$ , due to Proposition 2.2 along with Condition (i), that  $s_0 = o(\sqrt{n}/(\log p)^2)$ , and using Lemma 2.5. Similarly, and by symmetry, the second term goes to zero as  $n, p \rightarrow \infty$  and the claim follows.

We next focus on the event that  $t_0$ , given by (19) exists. By definition of  $t_0$  in this case, we have

$$\frac{pG(t_0)}{R(t_0) \vee 1} = q.$$

(Indeed, it is clear that the left-hand side is at most  $q$ . Equality holds since  $t_0$  is the minimum  $t$ , with such property.)

Define  $Q(t) \equiv G(t)/2$  for all  $t \in \mathbb{R}$ . Let

$$A_p \equiv \sup_{0 \leq t \leq t_p} \left| \frac{\sum_{i \in S_{\geq 0}} \{\mathbb{I}(T_i \leq -t) - Q(t)\} + \sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - Q(t)\}}{pG(t)} \right| \quad (44)$$

Then,

$$\begin{aligned}
\text{FDP}_{\text{dir}}(t_0) &= \frac{\sum_{i \in S_{\geq 0}} \mathbb{I}(T_i \leq -t_0) + \sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t_0)}{R(t_0) \vee 1} \\
&\leq \frac{pG(t_0)A_p + s_0Q(t_0) + 2(p - s_0)Q(t_0)}{R(t_0)} \quad (45)
\end{aligned}$$

$$\leq \frac{pG(t_0)(1 + A_p)}{R(t_0)} \leq q(1 + A_p). \quad (46)$$

Hence, we need to prove that  $A_p \rightarrow 0$ , in probability. Note that

$$A_p \leq \sup_{0 \leq t \leq t_p} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - Q(t)\}}{pG(t)} \right| + \left| \frac{\sum_{i \in S_{\geq 0}} \{\mathbb{I}(T_i \leq -t) - Q(t)\}}{pG(t)} \right| \right\} \quad (47)$$

$$\leq \sup_{0 \leq t \leq t_p} \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - Q(t)\}}{pG(t)} \right| \quad (48)$$

$$+ \sup_{0 \leq t \leq t_p} \left| \frac{\sum_{i \in S_{\geq 0}} \{\mathbb{I}(T_i \leq -t) - Q(t)\}}{pG(t)} \right|. \quad (49)$$

Note that by symmetry it is sufficient to prove that the first term in (47) goes to zero in probability. Consider a discretization  $0 \leq \tau_1 < \tau_2 < \dots < \tau_b = t_p$  such that  $\tau_j - \tau_{j-1} = v_p$ , for  $1 \leq j \leq b-1$  and  $\tau_b - \tau_{b-1} \leq v_p$ , where  $v_p = 1/\sqrt{\log p}$ . Hence,  $b \sim t_p/v_p$ . For any  $t \in [\tau_{j-1}, \tau_j]$ , we have

$$\begin{aligned} \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq \tau_j)}{pQ(\tau_j)} \cdot \frac{Q(\tau_j)}{Q(\tau_{j-1})} &\leq \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} \\ &\leq \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq \tau_{j-1})}{pQ(\tau_{j-1})} \cdot \frac{Q(\tau_{j-1})}{Q(\tau_j)} \end{aligned}$$

Hence, it suffices to show that

$$\max_{0 \leq j \leq b} \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq \tau_j) - Q(\tau_j)\}}{pQ(\tau_j)} \right| \rightarrow 0 \quad (50)$$

in probability.

In the following lemma, we provide sufficient conditions to obtain Eq. (50).

**Lemma 7.3.** Suppose that for any  $\delta > 0$ , the followings hold:

$$\sup_{0 \leq t \leq t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} = o(1) \quad (51)$$

and

$$\int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} dt = o(v_p) \quad (52)$$

where  $t_p = (2 \log p - 2 \log \log p)^{1/2}$  and  $v_p = (\log p)^{-1/2}$ , then (50) hold true.

We refer to Appendix A.2 for the proof of Lemma 7.3.

By virtue of Lemma 7.3 we only need to prove Eqs. (51) and (52). We start by analyzing the following expression

$$\begin{aligned}
\mathbb{E} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - \mathbb{P}(T_i \geq t)\}}{pQ(t)} \right|^2 \right\} & \quad (53) \\
& \leq \frac{\sum_{i,j \in S_{\leq 0}} \{\mathbb{P}(T_i \geq t, T_j \geq t) - \mathbb{P}(T_i \geq t)\mathbb{P}(T_j \geq t)\}}{p_0^2 Q(t)^2} \\
& \leq \frac{1}{p_0^2} \sum_{i,j \in S_{\leq 0}} \frac{\mathbb{P}(T_i \geq t, T_j \geq t)}{Q(t)^2} - 1 \\
& \leq \frac{1}{p_0^2} \sum_{i,j \in [p]} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} - 1, \quad (54)
\end{aligned}$$

with  $p_0 = |S_{\leq 0}|$  and

$$\tilde{T}_i \equiv \frac{\sigma}{\hat{\sigma}} \tilde{Z}_i + \frac{\Delta_i}{\hat{\sigma} \sqrt{\Lambda_{ii}}}. \quad (55)$$

The last inequality of (53) holds because  $\theta_{0,i} \leq 0$  for  $i \in S_{\leq 0}$  and therefore  $T_i \leq \tilde{T}_i$  (Recall definition of  $T_i$ , given by Eq. (39).) Further, because  $S^c = \{i \in [p] : \theta_{0,i} = 0\} \subseteq S_{\leq 0}$ , we have  $p_0 \geq p - s_0$ . Since  $s_0 = o(\sqrt{n}/(\log p)^2)$  by Condition (i), we have  $p_0 = \Omega(p)$ .

We partition the set  $\{(i, j) : i, j \in [p]\}$  into two disjoint sets, namely  $\Gamma(\gamma, c_0)$  (highly correlated test statistics) and  $\Gamma(\gamma, c_0)^c$  (weakly correlated test statistics). (Recall the definition of set  $\Gamma(\gamma, c_0)$  given by (21).) We analyze the contribution of each set separately.

### 7.1.1. Highly correlated test statistics ( $\Gamma(\gamma, c_0)$ )

We first consider the set  $\Gamma(\gamma, c_0)$ . Note that  $(\tilde{Z}_i, \tilde{Z}_j) \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \Lambda_{ij}^0 \\ \Lambda_{ij}^0 & 1 \end{bmatrix}\right)$ . Using (39), we have

$$\begin{aligned}
\mathbb{P}\left(\tilde{T}_i \geq t, \tilde{T}_j \geq t\right) & \leq \mathbb{P}\left(\tilde{Z}_i > \frac{\hat{\sigma}}{\sigma} t - \frac{\Delta_i}{\sigma \sqrt{\Lambda_{ii}}}, \tilde{Z}_j > \frac{\hat{\sigma}}{\sigma} t - \frac{\Delta_j}{\sigma \sqrt{\Lambda_{jj}}}\right) \\
& \leq \mathbb{P}\left(\tilde{Z}_i > (1 - \varepsilon_1)t - \varepsilon_2, \tilde{Z}_j > (1 - \varepsilon_1)t - \varepsilon_2\right) \\
& \quad + \mathbb{P}\left\{\|\Delta\|_{\infty} \geq \sigma \varepsilon_2 \sqrt{c}\right\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\} \quad (56) \\
& \leq C \left((1 - \varepsilon_1)t - \varepsilon_2 + 1\right)^{-2} \exp\left\{-\frac{((1 - \varepsilon_1)t - \varepsilon_2)^2}{1 + \Lambda_{ij}^0}\right\} \\
& \quad + \mathbb{P}\left\{\|\Delta\|_{\infty} \geq \sigma \varepsilon_2 \sqrt{c}\right\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\}, \quad (57)
\end{aligned}$$

where the last inequality follows from [Liu13, Lemma 6.2].

Let  $\Psi(\rho) \equiv \{(i, j) : i, j \in [p], |\Lambda_{ij}^0| > (1 - \rho)/(1 + \rho)\}$ . Note that by (41) and since  $|\Lambda^0 - \Omega^0|_\infty = o_p(1)$ , we have  $|\Psi(\rho)| = O(p)$ . We can write

$$\begin{aligned} \frac{1}{p_0^2} \sum_{(i,j) \in \Gamma(\gamma, c_0)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} &\leq \frac{1}{p_0^2} \left[ \sum_{(i,j) \in \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} \right. \\ &\quad \left. + \sum_{(i,j) \in \Gamma(\gamma, c_0) \setminus \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} \right]. \end{aligned} \quad (58)$$

We treat the terms on the right hand side separately. For the first term, since  $\Lambda_{ij}^0 \leq 1$ , by using (57), we have

$$\begin{aligned} \frac{1}{p_0^2} \sum_{(i,j) \in \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} &\leq \frac{|\Psi(\rho)|}{p_0^2 Q(t)^2} \left\{ C((1 - \varepsilon_1)t - \varepsilon_2)^{-2} \exp\left(-((1 - \varepsilon_1)t - \varepsilon_2)^2/2\right) \right. \\ &\quad \left. + \mathbb{P}\{\|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c}\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\} \right\} \\ &\leq Cp^{-1} \left\{ \left(\frac{G((1 - \varepsilon_1)t - \varepsilon_2)}{G(t)}\right)^2 \exp(((1 - \varepsilon_1)t - \varepsilon_2)^2/2) \right. \\ &\quad \left. + \frac{1}{Q(t)^2} \left(\mathbb{P}\{\|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c}\} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\}\right) \right\}, \end{aligned}$$

where in the second inequality we used the fact that  $|\Psi(\rho)| = O(p)$  and that  $p_0 = \Omega(p)$ . Take  $\varepsilon_2 = s_0(\log p)/\sqrt{n}$ . By using Lemmas 2.2, 2.5, and 7.2, we get

$$\begin{aligned} \frac{1}{p_0^2} \sum_{(i,j) \in \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} &\leq Cp^{-1} (1 + \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_2^2 t^2 + \varepsilon_1^2 t^4) e^{t^2/2} \\ &\quad + C \frac{p^{-1}}{Q(t)^2} \left( \mathbb{P}\{\|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c}\} \right. \end{aligned} \quad (59)$$

$$\begin{aligned} &\quad \left. + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\} \right) \\ &\leq Cp^{-1} \left\{ (1 + \varepsilon_1^2 + \varepsilon_2^2) e^{t^2/2} \right. \end{aligned} \quad (60)$$

$$\begin{aligned} &\quad \left. + e^{t^2/2} \varepsilon_2^2 t^2 + e^{t^2/2} \varepsilon_1^2 t^4 \right. \\ &\quad \left. + \frac{1}{Q(t)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P}\left\{\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \geq \varepsilon_1\right\} \right) \right\}, \end{aligned} \quad (61)$$

for some constant  $C > 0$ .

To bound the second term on the right-hand side of Eq. (58), note that for  $(i, j) \in \Gamma(\gamma, c_0) \setminus \Psi(\rho)$ , we have  $\Lambda_{ij}^0 \leq (1 - \rho)/(1 + \rho)$ . Thus, using (57)

$$\begin{aligned}
& \frac{1}{p_0^2} \sum_{(i,j) \in \Gamma(\gamma, c_0) \setminus \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} \\
& \leq \frac{|\Gamma(\gamma, c_0)|}{p_0^2 Q(t)^2} \left\{ C \left( (1 - \varepsilon_1)t - \varepsilon_2 \right)^{-2} \exp \left( -(1 + \rho) \left( (1 - \varepsilon_1)t - \varepsilon_2 \right)^2 / 2 \right) \right. \\
& \quad \left. + \mathbb{P} \left\{ \|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c} \right\} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right\} \\
& \leq c' p^{\rho-1} \left\{ \left( \frac{G \left( (1 - \varepsilon_1)t - \varepsilon_2 \right)}{G(t)} \right)^2 \exp \left( (1 - \rho) \left( (1 - \varepsilon_1)t - \varepsilon_2 \right)^2 / 2 \right) \right. \\
& \quad \left. + \frac{1}{Q(t)^2} \left( \mathbb{P} \left\{ \|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c} \right\} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \right\},
\end{aligned}$$

for any arbitrary constant  $c' > 0$ .

Hence, using Lemmas 2.2, 2.5, and 7.2, we get

$$\begin{aligned}
& \frac{1}{p_0^2} \sum_{(i,j) \in \Gamma(\gamma, c_0) \setminus \Psi(\rho)} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} \tag{62} \\
& \leq c' p^{\rho-1} \left( 1 + \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_2^2 t^2 + \varepsilon_1^2 t^4 \right) e^{(1-\rho)t^2/2} \\
& \quad + c' \frac{p^{\rho-1}}{Q(t)^2} \left( \mathbb{P} \left\{ \|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c} \right\} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \leq c' p^{\rho-1} \left\{ \left( 1 + \varepsilon_1^2 + \varepsilon_2^2 \right) e^{(1-\rho)t^2/2} + e^{(1-\rho)t^2/2} \varepsilon_2^2 t^2 + e^{(1-\rho)t^2/2} \varepsilon_1^2 t^4 \right. \\
& \quad \left. + \frac{1}{Q(t)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \right\} \tag{63}
\end{aligned}$$

uniformly for  $0 \leq t \leq t_p$ , and for any arbitrary constant  $c' > 0$ .

7.1.2. Weakly correlated test statistics  $(\Gamma(\gamma, c_0)^c)$ 

We next consider  $\Gamma(\gamma, c_0)^c \cap S_{\leq 0}$ . Using [Liu13, Lemma 6.1] (for  $d = 2$  in its statement) and Eq. (56), we have

$$\begin{aligned}
& \sup_{0 \leq t \leq t_p} \left| \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} - 1 \right| \\
& \leq \sup_{0 \leq t \leq t_p} \left| \frac{1}{Q(t)^2} \mathbb{P} \left( \tilde{Z}_i > \frac{\hat{\sigma}}{\sigma} t - \frac{\Delta_i}{\sigma \sqrt{\Lambda_{ii}}}, \tilde{Z}_j > \frac{\hat{\sigma}}{\sigma} t - \frac{\Delta_j}{\sigma \sqrt{\Lambda_{jj}}} \right) - 1 \right| \\
& \leq \sup_{0 \leq t \leq t_p} \left| \frac{1}{Q(t)^2} \mathbb{P} \left( \tilde{Z}_i > (1 - \varepsilon_1)t - \varepsilon_2, \tilde{Z}_j > (1 - \varepsilon_1)t - \varepsilon_2 \right) - 1 \right| \\
& + \frac{1}{Q(t_p)^2} \left( \mathbb{P} \{ \|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c} \} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \leq \sup_{0 \leq t \leq t_p} \left( \frac{Q((1 - \varepsilon_1)t - \varepsilon_2)}{Q(t)} \right)^2 \tag{64}
\end{aligned}$$

$$\begin{aligned}
& \sup_{0 \leq t \leq t_p} \left| \frac{\mathbb{P} \left( \tilde{Z}_i > (1 - \varepsilon_1)t - \varepsilon_2, \tilde{Z}_j > (1 - \varepsilon_1)t - \varepsilon_2 \right)}{Q((1 - \varepsilon_1)t - \varepsilon_2)^2} - 1 \right| \\
& + \sup_{0 \leq t \leq t_p} \left| \left( \frac{Q((1 - \varepsilon_1)t - \varepsilon_2)}{Q(t)} \right)^2 - 1 \right| \tag{65}
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{Q(t_p)^2} \left( \mathbb{P} \{ \|\Delta\|_\infty \geq \sigma \varepsilon_2 \sqrt{c} \} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \leq (1 + \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) C (\log p)^{-1 - \gamma_1} + C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) \\
& + \frac{1}{Q(t_p)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right), \tag{66}
\end{aligned}$$

for some constant  $C > 0$ , where  $\gamma_1 = \min(\gamma, 1/2)$ . In the last inequality above, we applied Lemma 7.1 (Note that  $Q(t) \equiv G(t)/2$  by definition). Therefore, by employing bound (66) for all  $(i, j) \in \Gamma(\gamma, c_0)^c$ , we get

$$\begin{aligned}
& \frac{1}{p_0^2} \sum_{(i,j) \in \Gamma(\gamma, c_0)^c} \frac{\mathbb{P}(\tilde{T}_i \geq t, \tilde{T}_j \geq t)}{Q(t)^2} - 1 \\
& \leq C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) (1 + (\log p)^{-1 - \gamma_1}) + C (\log p)^{-1 - \gamma_1} \\
& + \frac{1}{Q(t_p)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) + \frac{|\Gamma(\gamma, c_0)^c|}{p_0^2} - 1, \tag{67}
\end{aligned}$$

uniformly for  $0 \leq t \leq t_p$ , and for some positive constants  $C, c_1, c_2$ . Note that this inequality is obtained by applying .

Combining (53), (58) with bounds (61), (63) and (67), we obtain that

$$\begin{aligned}
& \mathbb{E} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - \mathbb{P}(T_i \geq t)\}}{pQ(t)} \right|^2 \right\} \\
& \leq C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) \left( 1 + (\log p)^{-1-\gamma_1} + p^{-1} e^{t^2/2} + p^{\rho-1} e^{(1-\rho)t^2/2} \right) \\
& \quad + \frac{1}{Q(t_p)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \quad + c' p^{\rho-1} e^{(1-\rho)t^2/2} + C p^{-1} e^{t^2/2} + C (\log p)^{-1-\gamma_1} + \frac{p^2}{p_0^2} - 1, \tag{68}
\end{aligned}$$

uniformly for  $0 \leq t \leq t_p$ , some positive constants  $C, c_1, c_2$  and for any constant  $c' > 0$ .

We are now ready to prove the conditions of Lemma 7.3, namely Eqs. (51) and (52). Fix arbitrary constant  $\delta > 0$ . By Chebyshev's inequality, we write

$$\begin{aligned}
& \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} \\
& \leq \frac{1}{\delta^2} \mathbb{E} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq t) - \mathbb{P}(T_i \geq t)\}}{pQ(t)} \right|^2 \right\} \\
& \leq \frac{1}{\delta^2} \left[ C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) \left( 1 + (\log p)^{-1-\gamma_1} + p^{-1} e^{t^2/2} + p^{\rho-1} e^{(1-\rho)t^2/2} \right) \right. \\
& \quad + \frac{1}{Q(t_p)^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \quad \left. + c' p^{\rho-1} e^{(1-\rho)t^2/2} + C p^{-1} e^{t^2/2} + C (\log p)^{-1-\gamma_1} + \frac{p^2}{p_0^2} - 1 \right], \tag{69}
\end{aligned}$$

where the second step follows from (68), uniformly for  $0 \leq t \leq t_p$  and for some constant  $C > 0$  and an arbitrarily small constant  $c' > 0$ . Hence, by substituting for  $t_p = (2 \log p - 2 \log \log p)^{1/2}$ , we obtain

$$\begin{aligned}
& \sup_{0 \leq t \leq t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} \\
& \leq \frac{1}{\delta^2} \left[ 4C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 (\log p)^2 + \varepsilon_2^2 \log p) \left( 1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-1+\rho} \right) \right. \\
& \quad + p^2 \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \quad \left. + C (\log p)^{-1} + c' (\log p)^{-(1-\rho)} + C (\log p)^{-1-\gamma_1} + \frac{p^2}{p_0^2} - 1 \right]. \tag{70}
\end{aligned}$$

Recall that  $\varepsilon_2 = s_0(\log p)/\sqrt{n}$ . We take  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$ . By [JM14a, Lemma 3.3], we have that for this choice of  $\varepsilon_1$ ,  $\mathbb{P} \{ |\hat{\sigma}/\sigma - 1| \geq \varepsilon_1 \} \rightarrow 0$  and hence Eq.(51) holds.

Likewise, (52) holds because continuing from (69) and by applying reverse Fatou Lemma, we can write

$$\int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pG(t)} - 1 \right| \geq \delta \right\} dt \leq \int_0^{t_p} \left[ C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 t_p^4 + \varepsilon_2^2 t_p^2) \left( 1 + (\log p)^{-1-\gamma_1} + p^{-1} e^{t^2/2} + p^{\rho-1} e^{(1-\rho)t^2/2} \right) + t_p^2 e^{t_p^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \right] dt \quad (71)$$

$$\leq C (\varepsilon_1^2 t_p + \varepsilon_2^2 t_p + \varepsilon_1^2 t_p^5 + \varepsilon_2^2 t_p^3) \left( 1 + (\log p)^{-1-\gamma_1} + p^{-1} e^{t^2/2} + p^{\rho-1} e^{(1-\rho)t^2/2} \right) + t_p^3 e^{t_p^2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \quad (72)$$

$$+ c' p^{\rho-1} t_p e^{(1-\rho)t_p^2/2} + C p^{-1} t_p e^{t_p^2/2} + C t_p (\log p)^{-1-\gamma_1} \leq 2C (\varepsilon_1^2 (\log p)^{5/2} + \varepsilon_2^2 (\log p)^{3/2}) \left( 1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-1+\rho} \right) + p^2 (\log p)^{-1/2} \left( e^{-c_1 n} + p^{-c_2} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \quad (73)$$

$$+ c (\log p)^{-(1/2-\rho)} + C (\log p)^{-1/2} + C (\log p)^{-1/2-\gamma_1} = o((\log p)^{-1/2}) = o(v_p). \quad (74)$$

In the last step we used the probabilistic bound on  $|\hat{\sigma}/\sigma - 1|$ , given in [SZ12, Theorem 2.1], with  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$ , and assumption  $s_0 = o(\sqrt{n}/(\log p)^2)$ . This shows that Eq. (52) holds and hence completes the proof.

## 7.2. Proof of Theorem 3.3

The threshold  $t_0$  returned by the FCD procedure is data-dependent. To analyze the power, we first upper bound  $t_0$  by a data-independent threshold  $t_*$ .

**Lemma 7.4.** Under the assumptions of Theorem 3.3, we have

$$t_0 \leq t_*, \quad t_* = \Phi^{-1} \left( 1 - \frac{qs_0}{2p} (1 - o(1)) \right).$$

Proof of Lemma 7.4 is given in Appendix A.3.

Since  $t_0 \leq t_*$  by Lemma 7.4, if we replace  $t_0$  by  $t_*$ , we obtain a lower bound on the power. For fixed arbitrarily small constants  $c_0, \delta, \varepsilon$ , define

$$\mathcal{G} = \mathcal{G}(\delta, c_0, \varepsilon) = \left\{ \max |\Lambda_{ii} - \Omega_{ii}| \leq c_0, \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \leq \delta, \|\Delta\|_\infty \leq \varepsilon \right\}.$$

Define  $S_+ \equiv \{i \in [p] : \theta_{0,i} > 0\}$  and  $S_- \equiv \{i \in [p] : \theta_{0,i} < 0\}$ . Therefore,  $S = S_+ \cup S_-$ . We have

$$\begin{aligned}
\text{Power} &= \mathbb{E} \left[ \frac{|\{j \in \widehat{S} : \widehat{\text{sign}}_j = \text{sign}(\theta_{0,j})\}|}{\max(|S|, 1)} \right] \\
&= \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P}(T_i \geq t_*) + \frac{1}{s_0} \sum_{i \in S_+^+} \mathbb{P}(T_i \leq -t_*) \\
&= \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P} \left( \frac{\sqrt{n} \widehat{\theta}_i^{\text{d}}}{\widehat{\sigma} \sqrt{\Lambda_{ii}}} \geq t_* \right) + \frac{1}{s_0} \sum_{i \in S_-} \mathbb{P} \left( \frac{\sqrt{n} \widehat{\theta}_i^{\text{d}}}{\widehat{\sigma} \sqrt{\Lambda_{ii}}} \leq -t_* \right) \\
&= \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P} \left( \frac{\sigma}{\widehat{\sigma}} \widetilde{Z}_i + \frac{\sqrt{n} \theta_{0,i} + \Delta_i}{\widehat{\sigma} \sqrt{\Lambda_{ii}}} \geq t_* \right) \tag{75}
\end{aligned}$$

$$+ \frac{1}{s_0} \sum_{i \in S_-} \mathbb{P} \left( \frac{\sigma}{\widehat{\sigma}} \widetilde{Z}_i + \frac{\sqrt{n} \theta_{0,i} + \Delta_i}{\widehat{\sigma} \sqrt{\Lambda_{ii}}} \leq -t_* \right) \tag{76}$$

Define  $\eta_i \equiv (\sqrt{n} \theta_{0,i} + \Delta_i) / (\sigma \sqrt{\Lambda_{ii}})$ . On event  $\mathcal{G}$ , we have

$$\eta_i \geq \eta_i^- \equiv \frac{\sqrt{n} \theta_{0,i} - \varepsilon}{\sigma \sqrt{\Omega_{ii} + c_0}}, \quad \eta_i \leq \eta_i^+ \equiv \frac{\sqrt{n} \theta_{0,i} + \varepsilon}{\sigma \sqrt{\Omega_{ii} - c_0}}.$$

Using Equation (76), we have

$$\text{Power} \geq \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P} \left( \left[ Z_i + \eta_i \geq \frac{\widehat{\sigma}}{\sigma} t_* \right] \cdot \mathbb{I}(\mathcal{G}) \right) \tag{77}$$

$$\begin{aligned}
&+ \frac{1}{s_0} \sum_{i \in S_-} \mathbb{P} \left( \left[ Z_i + \eta_i \leq \frac{-\widehat{\sigma}}{\sigma} t_* \right] \cdot \mathbb{I}(\mathcal{G}) \right) - \mathbb{P}(\mathcal{G}^c) \\
&\geq \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P} \left( \left[ Z_i + \eta_i^- \geq (1 + \delta) t_* \right] \cdot \mathbb{I}(\mathcal{G}) \right) \tag{78}
\end{aligned}$$

$$\begin{aligned}
&+ \frac{1}{s_0} \sum_{i \in S_-} \mathbb{P} \left( \left[ Z_i + \eta_i^+ \leq -(1 + \delta) t_* \right] \cdot \mathbb{I}(\mathcal{G}) \right) - \mathbb{P}(\mathcal{G}^c) \\
&= \frac{1}{s_0} \sum_{i \in S_+} \mathbb{P} (Z_i + \eta_i^- \geq (1 + \delta) t_*) \mathbb{P}(\mathcal{G}) \tag{79} \\
&+ \frac{1}{s_0} \sum_{i \in S_-} \mathbb{P} (Z_i + \eta_i^+ \leq -(1 + \delta) t_*) \mathbb{P}(\mathcal{G}) - \mathbb{P}(\mathcal{G}^c).
\end{aligned}$$

Recall that  $s_0 = o(\sqrt{n}/(\log p)^2)$  as per Condition (i), and by using Proposition 2.2 and lemma 2.5, event  $\mathcal{G}$  holds with high probability and indeed it is easy to see that for  $\theta_{\min} > (\sigma/\sqrt{n})\sqrt{2 \log(p/s_0)}$ , we have

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{P}(\mathcal{G}^c)}{1 - \beta(\theta_0, n)} = 0.$$

Therefore,

$$\liminf_{n \rightarrow \infty} \frac{\text{Power}}{1 - \beta(\theta_0, n)} \geq \liminf_{n \rightarrow \infty} \frac{1}{s_0(1 - \beta(\theta_0, n))} \left[ \sum_{i \in S_+} \mathbb{P}(Z_i + \eta_i^- \geq (1 + \delta)t_*) \right. \quad (80)$$

$$\left. + \sum_{i \in S_-} \mathbb{P}(Z_i + \eta_i^+ \leq -(1 + \delta)t_*) \right]. \quad (81)$$

Since the above bound holds for all  $\varepsilon, \delta, c_0 > 0$ , we get

$$\liminf_{n \rightarrow \infty} \frac{\text{Power}}{1 - \beta(\theta_0, n)} \geq \liminf_{n \rightarrow \infty} \frac{1}{s_0(1 - \beta(\theta_0, n))} \left[ \sum_{i \in S_+} \mathbb{P}\left(Z_i + \frac{\sqrt{n}\theta_{0,i}}{\sigma\sqrt{\Omega_{ii}}} \geq t_*\right) \right. \quad (82)$$

$$\left. + \sum_{i \in S_-} \mathbb{P}\left(Z_i + \frac{\sqrt{n}\theta_{0,i}}{\sigma\sqrt{\Omega_{ii}}} \leq -t_*\right) \right]$$

$$= \liminf_{n \rightarrow \infty} \frac{1}{s_0(1 - \beta(\theta_0, n))} \left\{ \sum_{i \in S} \left(1 - \Phi\left(t_* - \frac{\sqrt{n}|\theta_{0,i}|}{\sigma\sqrt{\Omega_{ii}}}\right)\right) \right\}$$

$$= \liminf_{n \rightarrow \infty} \frac{1}{(1 - \beta(\theta_0, n))} \left\{ \frac{1}{s_0} \sum_{i \in S} F\left(\frac{qs_0}{p}, \frac{\sqrt{n}|\theta_{0,i}|}{\sigma\sqrt{\Omega_{ii}}}\right) \right\} = 1. \quad (83)$$

The last step holds by using the definition of function  $F(\cdot, \cdot)$ , given by Equation (29), and the fact that  $Z_i|X \sim \mathbf{N}(0, 1)$ .

### 7.3. Proof of Theorem 4.1

The proof follows the proof of Theorem 3.1. Note that for the results of theorem to hold, it suffices that the conditions of Lemma 7.3 to be satisfied. The result in [JM18, Theorem 3.8], implies that under the conditions of Theorem 4.1, for some constants  $C, c$ , and  $n \geq \max(25 \log p, cs_0 \log(p/s_0))$ , we have

$$\mathbb{P}\left(\|\Delta\|_\infty \geq C\tau\sigma\sqrt{\frac{s_0}{n}} \log p\right) \leq 2pe^{-C_{\min}n/(16s_0)} + pe^{-n/1000} + 8p^{-1}. \quad (84)$$

Using this, under the assumptions of Theorem 4.1, letting  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n}$ , and following the same steps as in the proof of Theorem 3.1, we will reach the

following equation which is similar to Eq. (70)

$$\begin{aligned}
& \sup_{0 \leq t \leq t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} \\
& \leq \frac{1}{\delta^2} \left[ 4C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 (\log p)^2 + \varepsilon_2^2 \log p) (1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-1+\rho}) \right. \\
& \quad \left. + p^2 \left( 2pe^{-C_{\min} n / (16s_0)} + pe^{-n/1000} + 8p^{-1} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \right. \\
& \quad \left. + C (\log p)^{-1} + c' (\log p)^{-1+\rho} + C (\log p)^{-1-\gamma_1} + \frac{p^2}{p_0^2} - 1 \right]. \tag{85}
\end{aligned}$$

By taking  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$  in Eq. (70) and replacing  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n}$ , we deduce that Eq. (51) holds. Similarly, using Eq. (84), we reach

$$\begin{aligned}
& \int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pG(t)} - 1 \right| \geq \delta \right\} dt \\
& \leq 2C (\varepsilon_1^2 (\log p)^{5/2} + \varepsilon_2^2 (\log p)^{3/2}) \left( 1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-(1-\rho)} \right) \\
& \quad + p^2 (\log p)^{-1/2} \left( 2pe^{-C_{\min} n / (16s_0)} + pe^{-n/1000} + 8p^{-1} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\
& \quad + c (\log p)^{-(1/2-\rho)} + C (\log p)^{-1/2} + C (\log p)^{-1/2-\gamma_1}. \tag{86}
\end{aligned}$$

which is similar to Eq. (71). Again, by taking  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$  and  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n}$  we deduce that Eq. (52) holds too. Hence, the desired results hold under the conditions of the Theorem.

#### 7.4. Proof of Theorem 4.2

The proof is similar to the proof of Theorem 4.1. Here, using the result in [JM18, Theorem 3.13], under the conditions of Theorem 4.2, for some constants  $C$ ,  $c$ , and  $n \geq s_0 \log p$ , we have

$$\mathbb{P} \left( \|\Delta\|_{\infty} \geq C\tau\sigma\sqrt{\frac{s_0}{n}} \log p + C\sigma \min(s_0, s_{\Omega}) \frac{\log p}{\sqrt{n}} \right) \leq 2pe^{-C_{\min} n / (16s_0)} \tag{87}$$

$$+ pe^{-cn} + 8p^{-1}. \tag{88}$$

Here, by taking  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n} + \min(s_0, s_{\Omega}) \log p/\sqrt{n}$ , we will reach the following equation which is similar to Eqs. (70), (85)

$$\begin{aligned}
& \sup_{0 \leq t \leq t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} \\
& \leq \frac{1}{\delta^2} \left[ 4C (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1^2 (\log p)^2 + \varepsilon_2^2 \log p) (1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-1+\rho}) \right. \\
& \quad \left. + p^2 \left( 2pe^{-C_{\min} n / (16s_0)} + pe^{-cn} + 8p^{-1} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \right. \\
& \quad \left. + C (\log p)^{-1} + c' (\log p)^{-(1-\rho)} + C (\log p)^{-1-\gamma_1} + \frac{p^2}{p_0^2} - 1 \right].
\end{aligned}$$

By taking  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$  in Eq. (70) and replacing  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n} + \min(s_0, s_\Omega) \log p/\sqrt{n}$ , we deduce that Eq. (51) holds. Similarly, using Eq. (87), we reach

$$\begin{aligned} & \int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pG(t)} - 1 \right| \geq \delta \right\} dt \\ & \leq 2C \left( \varepsilon_1^2 (\log p)^{5/2} + \varepsilon_2^2 (\log p)^{3/2} \right) \left( 1 + (\log p)^{-1-\gamma_1} + (\log p)^{-1} + (\log p)^{-(1-\rho)} \right) \\ & \quad + p^2 (\log p)^{-1/2} \left( 2pe^{-C_{\min} n/(16s_0)} + pe^{-cn} + 8p^{-1} + \mathbb{P} \left\{ \left| \frac{\hat{\sigma}}{\sigma} - 1 \right| \geq \varepsilon_1 \right\} \right) \\ & \quad + c(\log p)^{-(1/2-\rho)} + C(\log p)^{-1/2} + C(\log p)^{-1/2-\gamma_1}. \end{aligned}$$

which is similar to Eqs. (71), (86). Again, by taking  $\varepsilon_1 = \sqrt{s_0(\log p)/n}$ ,  $\varepsilon_2 = (\log p)\tau_0\sqrt{s_0/n} + \min(s_0, s_\Omega) \log p/\sqrt{n}$  we deduce that Eq. (52) holds too. Hence, the desired results hold under the conditions of the Theorem.

### Acknowledgements

A. Javanmard was partially supported by the NSF CAREER Award 1844481 and a Google Faculty Research Award. A. Javanmard would also like to acknowledge the financial support of the Office of the Provost at the University of Southern California through the Zumberge Fund Individual Grant Program.

## Appendix A: Proof of Technical Lemmas

### A.1. Proof of Lemma 7.2

For  $t \geq 0$ , we write

$$\frac{G((1-\delta)t-\varepsilon)}{G(t)} = 1 + \frac{\int_{(1-\delta)t-\varepsilon}^t \phi(x) dx}{G(t)} \leq 1 + \frac{(\varepsilon + \delta t)\phi((1-\delta)t-\varepsilon)}{G(t)}, \quad (89)$$

where we used that  $\phi(t)$  is a decreasing function. We next separate the cases of  $t \in (0, 1)$  and  $t \geq 1$ .

For  $0 < t < 1$ , we use the following bound

$$\phi(t) \leq (\sqrt{4+t^2} - t)\phi(t) \leq G(t), \quad (90)$$

where the last step is due to Birnbaum [B<sup>+</sup>42].

Moreover, for all  $t \geq 0$ ,

$$\begin{aligned} \frac{\phi((1-\delta)t-\varepsilon)}{\phi(t)} &= \exp\left\{t(\delta t + \varepsilon) - \frac{1}{2}((1-\delta)t + \varepsilon)^2\right\} \\ &\leq \exp\left\{t(\delta t + \varepsilon)\right\} \leq e^2, \end{aligned} \quad (91)$$

because by our assumption  $\delta^2 t \leq 1$  and  $\varepsilon t \leq 1$ .

By employing Eqs. (90) and (91) into Eq. (89), we obtain

$$\frac{G((1-\delta)t-\varepsilon)}{G(t)} \leq 1 + e^2(\varepsilon + \delta t) \leq 1 + e^2(\varepsilon + \delta). \quad (92)$$

For  $t \geq 1$ , using Lemma 7.1, we have that  $G(t) \geq \phi(t)/t$  and hence by using Eq. (91) into Eq. (89), we get

$$\frac{G((1-\delta)t-\varepsilon)}{G(t)} \leq 1 + e^2 t(\varepsilon + \delta t). \quad (93)$$

The result follows by combining the bound (92) and (93).

### A.2. Proof of Lemma 7.3

We write

$$\begin{aligned}
& \mathbb{P} \left[ \max_{0 \leq j \leq b} \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq \tau_j) - Q(\tau_j)\}}{pQ(\tau_j)} \right| \geq \delta \right] \\
& \leq \sum_{j=1}^b \mathbb{P} \left[ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq \tau_j) - Q(\tau_j)\}}{pQ(\tau_j)} \right| \geq \delta \right] \\
& \leq \frac{1}{v_p} \int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} dt \\
& \quad + \sum_{j=b-1}^b \mathbb{P} \left[ \left| \frac{\sum_{i \in S_{\leq 0}} \{\mathbb{I}(T_i \geq \tau_j) - Q(\tau_j)\}}{pQ(\tau_j)} \right| \geq \delta \right]
\end{aligned}$$

Therefore, it suffices to show that

$$\int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} dt = o(v_p),$$

and

$$\sup_{0 \leq t \leq t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in S_{\leq 0}} \mathbb{I}(T_i \geq t)}{pQ(t)} - 1 \right| \geq \delta \right\} dt = o(1),$$

which are the conditions of the lemma.

### A.3. Proof of Lemma 7.4

We first show that  $t_* < \sqrt{2 \log(p/s_0)}$ . Assuming otherwise, we have  $G(t_*) < G(\sqrt{2 \log(p/s_0)})$  because  $G(t)$  is decreasing. By definition of  $t_*$ , and Lemma 7.1 this results in

$$\frac{qs_0}{p}(1 - o(1)) = G(t_*) < G(\sqrt{2 \log(p/s_0)}) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\log(p/s_0)}}{\sqrt{2 \log(p/s_0)}} \quad (94)$$

$$= \frac{s_0}{p\sqrt{\pi \log(p/s_0)}}, \quad (95)$$

which is a contradiction.

Now, given that  $t_* < \sqrt{2 \log(p/s_0)} < \sqrt{2 \log p}$ , if the claim is not true, by definition of  $t_0$ , we should have

$$\frac{2p(1 - \Phi(t_*))}{R(t_*) \vee 1} > q. \quad (96)$$

We next show that  $R(t_*) \geq s_0(1 - o(1))$ .

Define

$$\mathcal{G} = \mathcal{G}(\delta, c_0, \varepsilon) = \left\{ \max |\Lambda_{ii} - \Omega_{ii}| \leq c_0, |\hat{\sigma}/\sigma - 1| \leq \delta, \|\Delta\|_\infty \leq \varepsilon \right\}$$

Define  $\widehat{S}(t_*) = \{i \in [p] : |T_i| > t_*\}$ . Using Proposition (2.2), for fixed  $i \in S$ , we have

$$\begin{aligned} \mathbb{P}(i \notin \widehat{S}(t_*)) &= \mathbb{P}(|T_i| \leq t_*) \\ &= \mathbb{P}\left(\left|\frac{\sqrt{n}\theta_{0,i}}{\widehat{\sigma}\sqrt{\Lambda_{ii}}} + \frac{\sigma}{\widehat{\sigma}}Z_i + \frac{\Delta_i}{\widehat{\sigma}\sqrt{\Lambda_{ii}}}\right| \leq t_*\right), \end{aligned} \quad (97)$$

with  $Z_i \sim \mathbf{N}(0, 1)$ . Define  $\eta_i \equiv (\sqrt{n}\theta_{0,i} + \Delta_i)/(\sigma\sqrt{\Lambda_{ii}})$ . On event  $\mathcal{G}$ , we have

$$|\eta_i| \geq \eta_{i,*} \equiv \frac{\sqrt{n}|\theta_{0,i}| - \varepsilon}{\sigma\sqrt{\Omega_{ii} + c_0}}$$

Continuing from Equation (97), we have

$$\begin{aligned} \mathbb{P}(i \notin \widehat{S}(t_*)) &= \mathbb{P}\left(|Z_i + \eta_i| \leq \frac{\widehat{\sigma}}{\sigma}t_*\right) \leq \mathbb{P}\left(\left[|Z_i + \eta_{i,*}| \leq \frac{\widehat{\sigma}}{\sigma}t_*\right] \cdot \mathbb{I}(\mathcal{G})\right) + \mathbb{P}(\mathcal{G}^c) \\ &\leq \mathbb{P}\left(\left[\eta_{i,*} - \frac{\widehat{\sigma}}{\sigma}t_* \leq |Z_i|\right] \cdot \mathbb{I}(\mathcal{G})\right) + \mathbb{P}(\mathcal{G}^c). \end{aligned}$$

Given that  $\theta_{0,i} > (\sigma/\sqrt{n})\sqrt{2\Omega_{ii}\log(p/s_0)}$  and  $t_* < \sqrt{2\log(p/s_0)}$ , we can choose  $\delta, c_0, \varepsilon$  and  $\varepsilon_0$  small enough such that on event  $\mathcal{G} = \mathcal{G}(\delta, c_0, \varepsilon)$ ,

$$\eta_{i,*} - \frac{\widehat{\sigma}}{\sigma}t_* \geq t_*,$$

and therefore

$$\begin{aligned} \mathbb{P}(i \notin \widehat{S}(t_*)) &\leq \mathbb{P}\left(\left[\eta_{i,*} - \frac{\widehat{\sigma}}{\sigma}t_* \leq |Z_i|\right] \cdot \mathbb{I}(\mathcal{G})\right) + \mathbb{P}(\mathcal{G}^c) \\ &\leq \mathbb{P}((t_* \leq |Z_i|) \cdot \mathbb{I}(\mathcal{G})) + \mathbb{P}(\mathcal{G}^c) \\ &\leq G(t_*) + \mathbb{P}(\mathcal{G}^c) \\ &\leq \left(\frac{qs_0}{p}\right) + \mathbb{P}(\mathcal{G}^c) \end{aligned} \quad (98)$$

Since  $\mathbb{P}(\mathcal{G}^c) \rightarrow 0$  and  $s_0 = o(\sqrt{n}/(\log p)^2)$ , we can choose a deterministic sequence  $L_n \rightarrow \infty$ , arbitrarily slow, as  $n \rightarrow \infty$ , such that  $L_n\mathbb{P}(\mathcal{G}^c) \rightarrow 0$  and  $L_n(s_0/p) \rightarrow 0$ . Letting  $A_n \equiv (qs_0/p) + \mathbb{P}(\mathcal{G}^c)$ , we have  $L_n A_n \rightarrow 0$ .

By applying Markov inequality, we obtain

$$\begin{aligned} \mathbb{P}(|S \cap \widehat{S}(t_*)^c| \geq s_0 L_n A_n) &\leq \frac{1}{s_0 L_n A_n} \mathbb{E}(|S_0 \cap \widehat{S}(t_*)^c|) \\ &\leq \frac{s_0 A_n}{s_0 L_n A_n} = \frac{1}{L_n}, \end{aligned} \quad (99)$$

where the last inequality follows from (98). Therefore, with high probability,  $|S_0 \cap \widehat{S}(t_*)^c| \leq s_0 L_n A_n$ , which implies that

$$R(t_*) = |\widehat{S}(t_*)| \geq |S| - |S \cap \widehat{S}(t_*)^c| \geq s_0(1 - L_n A_n), \quad (100)$$

as claimed.

Now using Equation (100) in Equation (96), we arrive at

$$1 - \Phi(t_*) > \frac{qs_0}{2p}(1 - L_n A_n).$$

Therefore, for  $t_*$ , given by

$$t_* = \Phi^{-1}\left(1 - \frac{qs_0}{2p}(1 - 2L_n A_n)\right),$$

we reach a contradiction which proves our claim  $t_0 \leq t_*$  is correct. The proof is complete by noting that  $L_n A_n = o(1)$  by choice of sequence  $L_n$ .

#### A.4. Proof of Corollary 3.5

Define

$$\alpha_n = \frac{qs_0}{p}, \quad u_n \equiv \frac{\sqrt{n}\theta_{\min}}{\sigma\sqrt{\Omega_{ii}}}.$$

Using Corollary 3.4, it suffices to show that  $F(\alpha_n, u_n) = 1 - \Phi(\Phi^{-1}(1 - \alpha_n/2) - u_n) \rightarrow 1$ . Equivalently, we show that  $\Phi^{-1}(1 - \alpha_n/2) - u_n \rightarrow -\infty$ .

By Lemma 7.1, we have

$$G(\sqrt{2\log(2/\alpha_n)}) < \frac{2\phi(\sqrt{2\log(2/\alpha_n)})}{\sqrt{2\log(2/\alpha_n)}} < 2\phi(\sqrt{2\log(2/\alpha_n)}) = \alpha_n. \quad (101)$$

Since  $G$  is a decreasing function, by applying  $G^{-1}$  on both sides, we get

$$\Phi^{-1}(1 - \alpha/2) = G^{-1}(\alpha_n) \leq \sqrt{2\log(2/\alpha_n)}$$

Using the above bound, we have

$$u_n - \Phi^{-1}(1 - \alpha_n/2) > u_n - \sqrt{2\log(2/\alpha_n)} \quad (102)$$

By the assumption on  $\theta_{\min}$ , we have that the left-hand side of (102) goes to  $\infty$ , which completes the proof.

#### References

- [B<sup>+</sup>42] Z. W. Birnbaum et al. An inequality for mill's ratio. *The Annals of Mathematical Statistics*, 13(2):245–246, 1942.
- [BC13] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.
- [BC15] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

- [BC16] R. F. Barber and E. J. Candès. A knockoff filter for high-dimensional selective inference. *arXiv:1602.03574*, 2016.
- [BCC<sup>+</sup>18] A. Belloni, V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato. High-dimensional econometrics and generalized gmm. *arXiv preprint arXiv:1806.01888*, 2018.
- [BCH14] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.
- [BCS18] R. F. Barber, E. J. Candès, and R. J. Samworth. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.
- [BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- [BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *American Journal of Mathematics*, 37:1705–1732, 2009.
- [Büh12] P. Bühlmann. Statistical significance in high-dimensional linear models. *arXiv:1202.1377*, 2012.
- [BvdG11] P. Bühlmann and S. Van de Geer. *Statistics for high-dimensional data*. Springer-Verlag, 2011.
- [BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, pages 1165–1188, 2001.
- [CFJL18] E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [CP09] E. Candès and Y. Plan. Near-ideal model selection by  $\ell_1$  minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.
- [CT05] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.
- [FDLL17] Y. Fan, E. Demirkaya, G. Li, and J. Lv. Rank: large-scale inference with graphical nonlinear knockoffs. *arXiv preprint arXiv:1709.00092*, 2017.
- [FGH12] J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):1467–9868, 2012.
- [FHG12] J. Fan, X. Han, and W. Gu. Control of the false discovery rate under arbitrary covariance dependence (with discussion). *Journal of American Statistical Association*, 107:1019–1045, 2012.
- [FL01] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- [FL08] J. Fan and J. Lv. Sure independence screening for ultrahigh di-

- mensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
- [GLN02] C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.
- [GR04] E. Greenshtein and Y. Ritov. Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.
- [GT00] A. Gelman and F. Tuerlinckx. Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.
- [JM13a] A. Javanmard and A. Montanari. Model selection for high-dimensional regression under the generalized irrepresentability condition. In *Advances in Neural Information Processing Systems*, pages 3012–3020, 2013.
- [JM13b] A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1427–1434. IEEE, 2013.
- [JM14a] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
- [JM14b] A. Javanmard and A. Montanari. Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.
- [JM18] A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.
- [KF00] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, pages 1356–1378, 2000.
- [Liu13] W. Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, pages 2948–2978, 2013.
- [Lou08] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.
- [LS<sup>+</sup>14] W. Liu, Q.-M. Shao, et al. Phase transition and regularized bootstrap in large-scale  $t$ -tests with false discovery rate control. *The Annals of Statistics*, 42(5):2003–2025, 2014.
- [MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.
- [Owe05] A. B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):411–426, 2005.
- [RFZ<sup>+</sup>05] S.-Y. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Tay-

- lor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, et al. Hiv-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype b isolates and implications for drug-resistance surveillance. *The Journal of Infectious Diseases*, 192(3):456–465, 2005.
- [RTF16] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.
- [RTW<sup>+</sup>06] S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.
- [RYB03] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.
- [RZ11] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. 2011.
- [SBvdG10] N. Städler, P. Bühlmann, and S. Van de Geer.  $\ell_1$ -penalization for Mixture Regression Models (with discussion). *Test*, 19:209–285, 2010.
- [SRC<sup>+</sup>15] W. Sun, B. Reich, T. Cai, M. Guindani, and A. Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society, Series B*, 77:59–83, 2015.
- [SZ12] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.
- [Tib96] R. Tibshirani. Regression shrinkage and selection with the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.
- [Tuk91] J. W. Tukey. The philosophy of multiple comparisons. *Statistical Science*, pages 100–116, 1991.
- [vdGB09] S. Van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- [VdGBRD14] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
- [VdV00] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [Wai09] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.
- [Wu08] W. B. Wu. On false discovery control under dependence. *The Annals of Statistics*, 36:364–380, 2008.
- [XCML11] J. Xie, T. T. Cai, J. Maris, and H. Li. Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, 4(4):417–430, 2011.

- [Zha10] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.
- [ZY06] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.
- [ZZ11] C.-H. Zhang and S. Zhang. Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models. [arXiv:1110.2563](https://arxiv.org/abs/1110.2563), 2011.
- [ZZ14] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.